2017

A Statistical Framework for Improved Automatic Flaw Detection in Nondestructive Evaluation Images

Ye Tian
Iowa State University, tianye1984@gmail.com

Ranjan Maitra
Iowa State University, maitra@iastate.edu

William Q. Meeker
Iowa State University, wqmeeker@iastate.edu

Stephen D. Holland
Iowa State University, sdh4@iastate.edu

A Statistical Framework for Improved Automatic Flaw Detection in Nondestructive Evaluation Images

Abstract

Nondestructive evaluation (NDE) techniques are widely used to detect flaws in critical components of systems like aircraft engines, nuclear power plants and oil pipelines in order to prevent catastrophic events. Many modern NDE systems generate image data. In some applications an experienced inspector performs the tedious task of visually examining every image to provide accurate conclusions about the existence of flaws. This approach is labor-intensive and can cause misses due to operator ennui. Automated evaluation methods seek to eliminate human-factors variability and improve throughput. Simple methods based on peak amplitude in an image are sometimes employed and a trained-operator-controlled refinement that uses a dynamic threshold based on signal-to-noise ratio (SNR) has also been implemented. We develop an automated and optimized detection procedure that mimics these operations. The primary goal of our methodology is to reduce the number of images requiring expert visual evaluation by filtering out images that are overwhelmingly definitive on the existence or absence of a flaw. We use an appropriate model for the observed values of the SNR-detection criterion to estimate the probability of detection. Our methodology outperforms current methods in terms of its ability to detect flaws.

Keywords

Dynamic thresholding, Image processing, Matched filter, Noise-Interference Model, Probability of Detection, Signal-to-Noise Ratio

Disciplines

Aerospace Engineering | Statistics and Probability

Comments

This is an Accepted Manuscript of an article published by Taylor & Francis as Tian, Ye, Ranjan Maitra, William Q. Meeker, and Stephen D. Holland. "A Statistical Framework for Improved Automatic Flaw Detection in Nondestructive Evaluation Images." *Technometrics* 59, no. 2 (2017): 247-261. DOI: 10.1080/00401706.2016.1153000.

A Statistical Framework for Improved Automatic Flaw Detection in Nondestructive Evaluation Images

Ye Tian
Department of Statistics and Statistical Laboratory
Iowa State University
Ames, IA 50011
(tianye1984@gmail.com)

Ranjan Maitra
Department of Statistics and Statistical Laboratory
Iowa State University
Ames, IA 50011
(maitra@iastate.edu)

William Q. Meeker
Department of Statistics and the Center for Nondestructive Evaluation
Iowa State University
Ames, IA 50011
(wqmeeker@iastate.edu)

Stephen D. Holland
Department of Aerospace Engineering and Center for Nondestructive Evaluation
Iowa State University
Ames, IA 50011
(sdh4@iastate.edu)

Nondestructive evaluation (NDE) techniques are widely used to detect flaws in critical components of systems like aircraft engines, nuclear power plants and oil pipelines in order to prevent catastrophic events. Many modern NDE systems generate image data. In some applications an experienced inspector performs the tedious task of visually examining every image to provide accurate conclusions about the existence of flaws. This approach is labor-intensive and can cause misses due to operator ennui. Automated evaluation methods seek to eliminate human-factors variability and improve throughput. Simple methods based on peak amplitude in an image are sometimes employed and a trained-operator-controlled refinement that uses a dynamic threshold based on signal-to-noise ratio (SNR) has also been implemented. We develop an automated and optimized detection procedure that mimics these operations. The primary goal of our methodology is to reduce the number of images requiring expert visual evaluation by filtering out images that are overwhelmingly definitive on the existence or absence of a flaw. We use an appropriate model for the observed values of the SNR-detection criterion to estimate the probability of detection. Our methodology outperforms current methods in terms of its ability to detect flaws.

Keywords: Dynamic thresholding, Image processing, Matched filter, Noise-Interference Model, Probability of Detection, Signal-to-Noise Ratio

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

1 Introduction

Nondestructive evaluation (NDE) methods (Bray and Stanley, 1996; Shull, 2002; Heller, 2012) are used to examine and characterize materials and detect flaws in components without causing them irreversible. Different physical principles (Rummel, 1983; Silk et al., 1987; Bray and McBride, 1992) guide different NDE techniques, providing methods based on radiography (Halmshaw, 1982, 1991), ultrasound (Krautkramer and Krautkramer, 1990), eddy-currents (Kahn et al., 1977; Collins et al., 1985; Yang et al., 2010), radiology (Martz et al., 2002), active thermography (Spicer and Osiander, 2002), acoustic emissions (Prosser, 2002), magnetic particles (Lindgren et al., 2002), liquid penetrants (Halmshaw, 1991) and other techniques (Shull, 2002; Heller, 2012). One of the primary purposes of NDE is to detect flaws in critical system components. Examples include fatigue cracks in aircraft engine turbine disks or blades, material anomalies in billets or forging materials that can be detected during manufacturing processes. Flaw detection is important in almost all cases but especially when there is a risk of such flaws causing serious or disastrous damage to systems (e.g., aircraft or bridges). Characterization of flaws after detection is also important in the application of nondestructive inspection, because it helps maintenance engineers obtain knowledge about flaw types, shape, size, location, and orientation, and use this information to decide on whether a specific part should continue in service or be immediately repaired or replaced.

There is substantial and long-standing interest in the development of statistical methodology and algorithms for the analysis of NDE data (*e.g.*, see Berens and Hovey, 1981, 1982, 1983, 1984; Gray and Thompson, 1986; Annis and Erland, 1989; Burkel et al., 1996; Hovey and Berens, 1988; Perdijon, 1988a,b, 1989; Neal and Speckman, 1993; Howard and Gilmore, 1994; Sweeting, 1995; Spencer and Schurman, 1995; Olin and Meeker, 1996; Howard et al., 1998; Aoki and Suga, 1999; Zaki et al., 2001; Legendre et al., 2001; Meyer and Candy, 2002; Zavaljevski et al., 2005; Dogandzic and Zhang, 2007; Hasanzadeh et al., 2008; Li and Meeker, 2009; Li et al., 2010; Gao and Meeker, 2012; Ng et al., 2013). Much of the NDE literature has revolved around the issues of noise reduction and comcomitant increased signal-to-noise ratio (SNR) and development of methodologies (*e.g.*, development of wavelet methods, expectation-maximization algorithm-type methods) for better estimating the extent and probability of detection (POD) of a flaw in different techniques. Some other attempts have been in the area of automated inspections. A comprehensive review of statistical issues and development in NDE is provided in the discussion paper of Olin and Meeker (1996). Another review of available methodology and techniques in this research area is provided in MIL-HDBK-1823A (2009).

With the rapid advances in technology, there has also been a quantum jump in the development of NDE techniques. More modern and automated methods of data acquisition have also resulted in

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

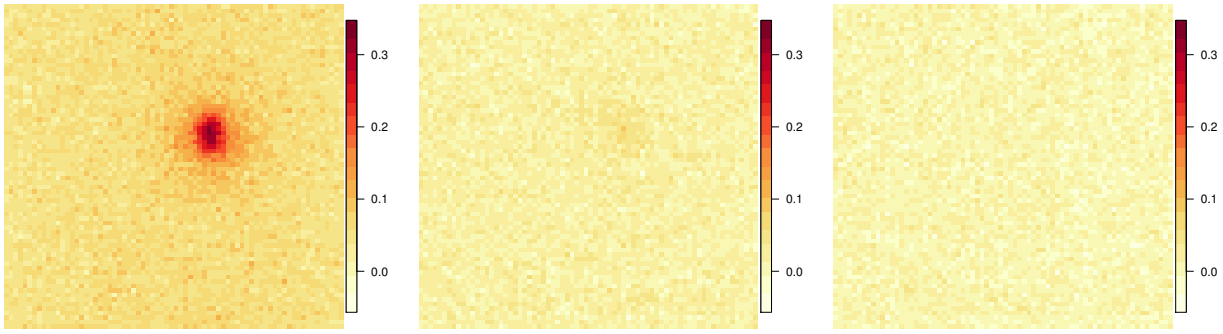
the capability of obtaining high-quality image-based data. Such datasets have obvious advantages in that images are natural objects for a technician to examine and use in order to make decisions. It is often much easier and more straightforward to detect the existence of a flaw or to assess flaw characteristics (*e.g.*, size and orientation) from visual inspection of an image than from a series of numbers provided by traditional methods. Image data, however, also provide challenges in terms of interpretation and detection as illustrated next in the context of vibrothermography which also forms the showcase application of this paper.

1.1 Analysis of Vibrothermography Image Data

Vibrothermography – also called sonic infrared, thermoacoustics or thermosonics – is a modern NDE imaging technique (Maldague, 2001; Henneke and Jones, 1979; Reifsnider et al., 1980; Holland, 2007) for detecting cracks or flaws in industrial, dental and aerospace applications. The imaging modality works on the principle that a sonic or ultrasonic energy pulse when applied to a unit causes it to vibrate. As a result, it is expected that the faces of a crack will rub against each other, resulting in an increase in temperature in that region. An infrared camera captures this increased temperature and produces a sequence of images of the temperature intensities over a short period of time starting just before the pulse of energy is applied and ending around the time that generated heat has dissipated. A sequence of images records the temperature changes over time. The primary objective of this technology is to detect flaws in the material with high precision. If the crack is larger than a certain threshold, the part needs to be repaired or replaced. Another goal is to predict the progression of the flaw whose sizes are below a certain threshold and therefore not cause for immediate concern, but important enough to suggest a purposeful schedule for future inspections.

Although the vibrothermography technology is still in its infancy, especially when compared to NDE images obtained using ultrasonic or radiographic methods, a commonly-used pre-processing data reduction technique (essentially eliminating the temporal dimension) is to use the image frame with the largest contrast (highest signal) (Li et al., 2010, 2011; Gao and Meeker, 2012) before analysis. In this paper, we consider these summary thermal images as the starting point for our methodological development and analyses. Figure 1 displays three sample thermal images, each obtained from a vibrothermographic time-course sequence of 150 image frames collected on a titanium Ti-6Al-4V specimens with known flaw sizes (if present). Figures 1a and b display strong and weak signals in the thermal image as a result of a larger- and smaller-sized flaw while Figure 1c is a thermal image of a specimen with no flaw and thus is essentially an image only of the noise in image acquisition. The three cases in Figure 1 illustrate the challenges in determining the presence and size of a flaw and if action is needed in terms of repair or replacement of the associated part. In Figure 1a, the flaw can

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.



(a) Strong signal (b) Weak signal (c) Noise image

Figure 1: Vibrothermography images on Ti-6Al-4V specimens with (a) a single large flaw, (b) a single small flaw and (c) no flaws.

be easily identified by a trained operator. Distinguishing a case with a weak signal from one that only images noise is more challenging, however. Similar challenges also exist with respect to other NDE imaging techniques such as those involving eddy currents (Hasanzadeh et al., 2008), ultrasound (see, *e.g.*, Legendre et al., 2001; Ng et al., 2013) and radiography (Wang and Liao, 2002).

1.2 Background and Current Practices

The development of analytical tools for vibrothermography image data is still in an early phase, so we discuss current practices in the context of other NDE imaging techniques, while noting that some of these practices have also carried over to vibrothermography. A common approach, largely a cross-over from analysis methods used in conventional ultrasonic NDE imaging, is to use the peak-amplitude method to detect a "hot spot" or region of elevated temperature readings (often after some initial signal processing of the image). Here, an operator visually finds the hottest spot in an image and uses the pixel intensity as the response for that image. Statistical models are then fit to these responses and the result is used to claim detection of a flaw when the observed response is higher than some threshold.

A more refined approach (Howard and Gilmore, 1994) to the above is used in the analysis of multi-zone ultrasound NDE images. Here, an identified hotspot in a processed image is enclosed within two visually-drawn rectangles in such a way that the pixels inside the inner rectangle have elevated intensities while the region enclosed between the smaller and the larger rectangles has almost no high-intensity pixels. Thus, the inner rectangle conceptually encloses the signal region while the outer (rectangular) frame represents the noise measurements. The pixels in these two regions are used to calculate the SNR from which it is determined, as before, if the region enclosing the hotspot is higher than the threshold, pointing to a possible flaw.

There has been some attention paid to the issue of thresholding in NDE imaging. Howard et al. (1998) proposed an SNR-based dynamic algorithm for detecting flaws in ultrasonic C-scan images. Jansohn and Schickert (1998) described a statistical threshold detection algorithm to interpret C-scan

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

images of concrete elements by modeling the signal amplitude with a lognormal or, alternatively, a Weibull distribution. Other work has focused on improving the quality of image processing. For instance, Chen and Wang (2004) utilized independent components analysis (ICA) for reducing speckle and to enhance edges of ultrasonic C-scan images. Li et al. (2010) introduced a three-dimensional (3-D) matched filter to enhance the SNR of the vibrothermography image sequence and also statistical methods for flaw detection using a matched filter output. Gao and Meeker (2012) presented methodology for the systematic analysis of image data from vibrothermography inspections, based on principal components and robust regression.

An important issue in all methods, not addressed in the literature, is the reliance on the human operator to manually identify, after the initial processing, the hotspot and, for Howard and Gilmore (1994)'s approach, to also draw the two rectangles. While the human eye and the visual system have an unmatched ability in detecting and resolving many situations, it is also true that operator fatigue and the low probability of finding a flaw in most NDE inspections greatly impacts detection accuracy, resulting in increased potential for missing an actual flaw. Thus, an approach which reduces this potential for human-factor misses would be desirable. In this paper, we develop a automated statistical algorithm that can identify images containing some evidence of the existence of a flaw. The objective behind this algorithm is to identify images where the existence or otherwise of a flaw is easily and conclusively established with a view to screening out images that require no further evaluation by a trained expert. This will reduce the volume of images required to be processed manually by a human inspector and potentially decrease the probability of human-factor misses.

As pointed out by the Editor, different aspects of the issue of flaw detection in NDE images have similarities with the image processing techniques of edge detection (Marr and Hildreth, 1980; Rosenfeld, 1984; Bergholm, 1987; Gauch and Pizer, 1993; O'Sullivan and Qian, 1994) and image segmentation (Rosenfeld and Kak, 1982; Qiu and Sun, 2007, 2009). Qiu (2005) points out that many of these techniques can be framed in the context of jump regression. In edge detection, the objective is to locate an object or object(s) in an image. Although many methods are available, O'Sullivan and Qian (1994) provided an edge detection algorithm using the *contrast statistic* to detect images of single objects of arbitrary shape and size. For multiple objects, they suggested iteratively locating boundaries one at a time. The contrast statistic was applied to both emission computed tomography image data as well as to simulations from a digitized phantom experiment. O'Sullivan and Qian (1994) provided generalizations to boundary detection in multi-channel and volumetric images, but the performance of these methods in situations involving only noise (such as would arise in most routine NDE inspections) is unclear. Image segmentation forms a major sub-area within the ambit of image processing, with the goal being to partition an image into different regions, each of which contain pixels that are similar to

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

others (with respect to some characteristic) in the same region, but different in that same characteristic from pixels in another region. Examples of such characteristics are the biochemical status (*e.g.*, in emission tomography) or composition (*e.g.*, in Magnetic Resonance Imaging) of tissue or land types in images obtained by remote sensing. Our application in NDE imaging is, however, centered solely on identifying possible hotspots in an image (in actual NDE applications, the occurrence of a flaw is a rare event), and then determining whether such a hotspot (if identified) is a flaw or simply an artifact of noise. In this paper therefore, we develop an automated statistical technique to identify potential hotspots in NDE images.

1.3 Overview

The remainder of this paper is organized as follows. Section 2 develops our automated feature detection algorithm and also incorporates extensions to account for real-life special cases. We also develop detection criterion and modeling strategies relating the flaw size and NDE metrics extracted from our proposed automated approach. The performance of the proposed methodology is evaluated on vibrothermography and simulated datasets in Section 3. The paper concludes with some discussion in Section 4. This paper also has an online supplement providing additional details on experimental illustrations, performance evaluations, and data analysis. Sections and figures in the supplement referred to in this paper are labeled with the prefix "S-".

2 Methodology

2.1 Preliminaries

Let $\boldsymbol{Z} \equiv \{Z(u,v) : u = 1, 2, \ldots, n_1; v = 1, 2, \ldots, n_2\}$ denote the observed $n_1 \times n_2$ image with $Z(u,v)$ as the observed intensity at the pixel with coordinates (u,v). Let us denote the true image signature using $\boldsymbol{\tau} \equiv \{\tau(u,v) : u = 1, 2, \ldots, n_1, v = 1, 2, \ldots, n_2\}$ and the noise with $\boldsymbol{\epsilon} \equiv \{\epsilon(u,v) : u = 1, 2, \ldots, n_1, v = 1, 2, \ldots, n_2\}$. Further, let μ be the systematic error at each pixel coordinate. Then the model is given by

$$\boldsymbol{Z} = \mu \mathbf{1} + \boldsymbol{\tau} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{1}$ is the vector that takes the value 1 for all coordinates and $\boldsymbol{\epsilon}$ is distributed according to some multivariate density $f(\cdot)$ with center zero in each coordinate.

We also review a few common NDE terms used in this paper. A *crack*, *inclusion* or *porosity* is the name given to different kinds of flaws that can affect the functioning or longevity of a system component or part. The term *flaw* itself is generically used to describe cracks, inclusions or porosities. Flaws are often seeded into a test block or individual specimens that will be used in experiments for evaluating performance of inspection systems; such seeded flaws are called *targets* because their existence

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

and location is known to the experimenters (but generally not to the inspectors used in experiments). Finally, the term *indication* is used to denote a hotspot, whether caused by a signal emanating from a true flaw or a noise artifact.

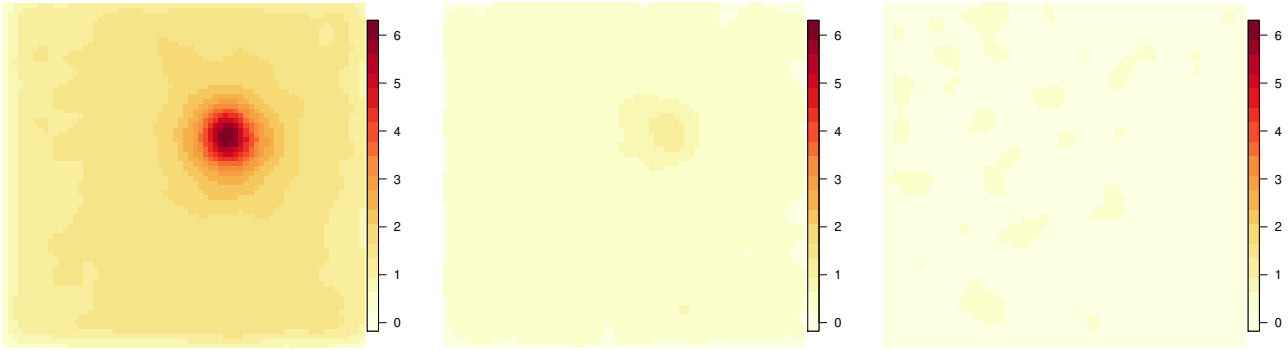
2.1.1 Image Processing with Matched Filters

Image processing methods are used to improve SNR. One approach, suitable for vibrothermography image data, is the matched filter (Turin, 1960, 1976) which we review here. Mathematically, given the model (1), the matched filter output is, notionally, the two-dimensional (2-D) convolution of the true signature $\boldsymbol{\tau}$ with the observed $\boldsymbol{Z}$ (*i.e.*, $\hat{\boldsymbol{\tau}} = \boldsymbol{\tau} \star \boldsymbol{Z}$, with $\star$ denoting the convolution operator). Note here that the true signature is not known (and is desired to be estimated), so in practice, the matched filter is implemented by convolving the known (or approximate) signature (also called a *template*) with the observed data (consisting of signal plus noise) to detect the presence of the signature in the output. A matched filter is an optimal linear filter in the sense that it maximizes the SNR in the presence of stationary white noise when a signal's impulse response function (or signature) is known. Even with spatially-correlated noise, the matched filter generally performs well. For more details, see, for example, Turin (1960), Turin (1976) or Engelberg (2007, Chapter 6).

Matched filtering has its origins in communications and signal processing and has been used as an image processing tool for various NDE imaging methods. For example, when the approximate signal signature is known, Li et al. (2010) used a 3-D matched filter to process the sequence-of-image data from vibrothermography experiments for titanium Ti-6Al-4V specimens containing fatigue cracks. Also, in ultrasonic inspections, the high background noise in titanium-alloy parts makes detection difficult, especially for small flaws. Matched filter processing has the ability to signficantly enhance SNR and improve the probability of flaw detection. While there are several ways to construct a matched filter, one approach that we have found promising utilizes the Gaussian shape of the flaw signature. The basic idea – modified from Li et al. (2010) to apply to a 2-D setting – places a radially symmetric 2-D Gaussian kernel at the center of the imaging region. At the pixel (u, v), this (discrete) Gaussian signature is specified by $f(u, v) \propto \exp\left[-2\sigma_{\hbar}^{-2}\{(u-\bar{u})^2 + (v-\bar{v})^2\}\right]$, where the center of the imaging region is at $(\bar{u}, \bar{v})$ and the filter bandwidth is defined in terms of its full-width-at-half-maximum (FWHM) of $\hbar$ pixels. (In signal and image processing parlance, the FWHM of a filter is the range of the interval formed by the two points where the signal attains half its peak value. For a Gaussian filter with scale parameter $\sigma_{\hbar}$, this translates to $\hbar = 2.355\sigma_{\hbar}$.) The choice of $\hbar$ is determined collectively by two criteria: (a) the Gaussian profile is such that it vanishes on the boundary pixels and (b) the profile provides reasonable resolution of the filtered output image in the sense that is a moderate proportion of non-negligible-valued pixels in the imaging grid.

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

In the specific experiments reported in this paper, we have a imaging grid of 30×30-pixels. The Gaussian shape used here is motivated by the thermal Green's function (Beck et al., 1992) which demonstrates that the temperature profile of an impulse point-source is of Gaussian form. The actual profiles in vibrothermography experiments vary from the Gaussian form because the geometric shape of the heat source comes from the locations of heating in the crack, and because the temperature profile is stepped rather than an impulse (Holland, 2011). Nevertheless, the diffusive nature of the heat conduction equation means that the Gaussian shape provides a good approximation. For our experiments,



(a) Strong signal (b) Weak signal (c) Noise image

Figure 2: Matched filter output for the vibrothermography image data of Figure 1

we have found $\hbar = 4.71$ (equivalently $\sigma_\hbar = 2$) to satisfy the criteria (a) and (b), though any similar value could also have been used. Figure 2 illustrates the results of matched filtering on the data of Figure 1. There is substantial noise reduction (and SNR enhancement) from the corresponding images of Figure 1. However, it also presents challenges, especially with regard to distinguishing material with a small flaw (Figure 2b) from a flawless specimen (Figure 2c). Such processed images (of $\hat{\boldsymbol{\tau}}$, which for notational convenience, we henceforth denote as $\boldsymbol{Y} = \{Y(u,v) : u = 1, 2, \ldots, n_1; v = 1, 2, \ldots, n_2\}$) form the input for our proposed automatic statistical screening and flaw detection algorithm.

We conclude discussion here by noting that Wiener filtering (Jain, 1989; Engelberg, 2007; Gonzalez and Woods, 2008) is often also used in NDE imaging instead of matched filtering. In this paper however, we focus development on matched-filtered images because this filter has optimality properties with respect to SNR and because our feature extraction and statistical modeling procedures are SNR-based. However, our development is also generally applicable to Wiener-filtered images.

2.1.2 SNR Computations in NDE Applications

The traditional definition of SNR in NDE applications compares the ratio of the peak signal intensity to that of peak noise (after removing the estimated bias, *i.e.* the effect of μ in (1), for both terms). Though there are generally many ways of defining SNR, we adopt this definition for consistency with the application in this paper. Specifically, SNR $= (\check{Y} - \bar{e})/(\check{e} - \bar{e})$, where $\check{Y}$ is the maximum (peak) value of the estimated signal (or hotspot value) of the processed image while $\check{e}$ and $\bar{e}$ are the peak

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

and average values of the estimated noise in the pixels. We now adapt Howard and Gilmore (1994)'s general strategy of identifying regions having signal and noise (the inner rectangle and the outer frame for their specific approach) to determine estimates for $\check{Y}$, $\check{e}$, or $\bar{e}$.

2.2 Automated Feature Extraction Algorithm

This section develops methodology to automate the process of flaw detection for NDE images, with special reference to vibrothermography image data. Our starting point is the operator-controlled rectangle-drawing approach of Howard and Gilmore (1994). However, visual examination of large collections of vibrothermography and ultrasonic inspection images of specimens containing flaws leads us to propose that the shape of most flaw signals can be described by ellipses. Thus, although one could adapt nonparametric edge-detection methods (O'Sullivan and Qian, 1994; Qiu, 2005), we choose to use a defined parametric shape in order to provide more power to our detection algorithm, especially in low SNR situations. Therefore, we assume that the signal region is elliptically-shaped. Generally a flaw signal can be covered by an ellipse which we call the "inner ellipse." We then draw another "outer ellipse" with identical center and orientation as its inner counterpart where the region between the two ellipses is expected to contain noise pixels only. (These ellipses are elliptical cousins of Howard and Gilmore (1994)'s inner and outer rectangles). We desire the inner ellipse to be as compact as possible, so that it not only covers most of the signal pixels, but also few or no noise pixels. After determining the location and orientation of the inner ellipse, we use an equal-area constraint (*i.e.*, the inner ellipse and the frame have the same area) to determine the outer ellipse.

The use of the outer ellipse adjacent to the inner ellipse (which is designed to encase the signal region) is because the noise level can vary considerably within the inspected specimen. Allowing the outer ellipse to have twice the area of its inner counterpart ensures equal number of pixels in both the inner ellipse and the adjoining frame, providing balance in precision. However, it is certainly possible that a few pixels in the frame have contamination from the signal, so we need to choose the optimal elliptical regions to provide the best contrast (SNR) for detection of a true flaw.

2.2.1 Optimal elliptical regions

As mentioned above, we want the inner ellipse to be compact, containing as many signal pixels and as few noise pixels as possible. Although human inspectors can manually draw such ellipses, we aim to provide an automated procedure that will increase throughput and reduce human factors variability. We formulate the choice of the inner ellipse in terms of an optimization problem, starting first with the case of only one indication in the image. Then the main steps of our algorithm are:

1. *Determine the center of the signal:* Assume that the ellipse is centered at the "hottest spot" (pixel with highest intensity $\check{Y}$) having coordinates $(u_\bullet, v_\bullet)$. Denote the ellipse with parameters (a, b, θ)

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

as $\mathcal{E}_i(a, b, \theta)$, where a and b are one-half of the lengths of the major and minor axes and θ is the angle between the major and the horizontal axes.

2. *Determine the inner ellipse*: We introduce the concept of ellipse "volume" $\mathcal{V}(a, b.\theta)$ as a function of (a, b, θ) as follows: For a given $\mathcal{E}_i(a, b, \theta)$, we define $\boldsymbol{Y}^i = \{Y(u, v) : (u, v) \in \mathcal{E}_i(a, b, \theta)\}$ and let $\boldsymbol{Y}^o = \boldsymbol{Y} \setminus \boldsymbol{Y}^i$ be the set of values $Y(u, v)$ outside the ellipse $\mathcal{E}_i(a, b, \theta)$. Further, let $\hat{\mu}$ be an estimate of μ. Intuitively, we set $\hat{\mu}$ to be equal to the mean of the intensities in $\boldsymbol{Y}^o$ (*i.e.*, the mean pixel intensity outside $\mathcal{E}_i(a, b, \theta)$). We subtract $\hat{\mu}$ from each element in $\boldsymbol{Y}$ to yield "corrected" intensities $\boldsymbol{Y}_c = \{Y_c(u, v)\}$, where $Y_c(u, v) = Y(u, v) - \hat{\mu}$. (Let $\boldsymbol{Y}_c^i$ and $\boldsymbol{Y}_c^o$ be the corresponding corrected intensities for pixels inside and outside the ellipse.) Intuitively, almost all of $\boldsymbol{Y}_c^i$ will have positive values, but the surrounding (contaminating) noise pixel intensities (in $\boldsymbol{Y}_c^i$) will be scattered around zero (taking either positive or negative values). Then the volume is defined as

$$\begin{aligned} \mathcal{V}(a, b, \theta) &= \sum_{(u,v)\in\mathcal{E}_i(a,b,\theta)} Y_c^i(u, v)\mathcal{I}_{[Y_c^i(u,v)>0]} + \lambda \sum_{(u,v)\in\mathcal{E}_i(a,b,\theta)} Y_c^i(u, v)\mathcal{I}_{[Y_c^i(u,v)\leq 0]} \\ &\equiv \sum_{(u,v)\in\mathcal{E}_i(a,b,\theta)} Y_c^i(u, v) + (\lambda - 1) \sum_{(u,v)\in\mathcal{E}_i(a,b,\theta)} Y_c^i(u, v)\mathcal{I}_{[Y_c^i(u,v)\leq 0]} \end{aligned} \tag{2}$$

   where $\mathcal{I}_{[\cdot]}$ is the indicator function and λ is a regularization parameter that can be used to control the compactness of $\mathcal{E}_i(a, b, \theta)$. For a fixed λ, the volume is a function of (a, b, θ). Intuitively again, a large volume (*i.e.*, a large $\mathcal{V}(a, b, \theta)$ value) corresponds to a compact inner ellipse containing mostly signal pixels and only a small number of noise pixels. Therefore our goal reduces to maximizing the volume $\mathcal{V}(a, b, \theta)$ as a function of (a, b, θ). The exact choice of λ is left as a control parameter, depending on the application, and the distinguishability of the true flaw signal $\tau(u, v)$ relative to the standard error in the image.

3. *Drawing the outer ellipse:* Once the inner ellipse is drawn in Step 2 , the outer ellipse is drawn using the same center and orientation (as that of the inner ellipse) but having twice the area (equivalently, the annular portion of the outer ellipse has the same area as the inner ellipse). Within this framework, the outer ellipse is specified to have axes lengths $a^* = a + \Delta$ and $b^* = b + \Delta$. From the area restriction, we have $a^*b^* = 2ab$ so that $\Delta = [-(a + b) + \sqrt{a^2 + b^2 + 6ab}]/2$. The outer ellipse thus drawn is used in obtaining the $\check{e}$ and $\bar{e}$ for our SNR calculations.

Our volume criterion has similarities with the contrast statistic in edge detection of O'Sullivan and Qian (1994). We now provide some theoretical basis for the selection of λ and for the use of $\mathcal{V}(a, b, \theta)$.

Theoretical justification of the volume criterion and guidance for selecting λ: Suppose that the true signal region is indeed elliptically-shaped, with all positive pixel intensities inside the ellipse (*i.e.*, $\tau(u, v) > 0$ for $(u, v) \in \mathcal{E}_i(a, b, \theta)$ and zero otherwise). Suppose also, for simplicity, that there is only

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

one signal region in the image, and that $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ is modeled in terms of Gaussian white noise, that is, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \varsigma^2 \boldsymbol{I})$, where $\mathbf{0}$ is the vector having zeroes in all coordinates and $\boldsymbol{I}$ is the identity matrix of appropriate order. Given $\mathcal{E}_i(a, b, \theta)$, one may use $\hat{\mu} = \bar{Y}^o = \mathbf{1}'\boldsymbol{Y}^o/n_o$ to estimate μ. In our implementation, we compared results using the sample mean $\bar{Y}^o$ as well as the sample median $\tilde{Y}^o$ with virtually indistinguishable performance so we only report results using $\bar{Y}^o$. In any case, both $\bar{Y}^o$ and $\tilde{Y}^o$ are consistent estimators for μ and since n_o, the number of pixels outside $\mathcal{E}_i(a, b, \theta)$ is anticipated to be large, large-sample arguments hold. Then, $\boldsymbol{Y}_c^i$ and $\boldsymbol{Y}_c^o$ are the corrected intensities inside and outside the ellipse after accounting for the systematic estimated bias. In particular, in terms of n_o, each $Y_c^i(u, v)$ is normally distributed with mean $\tau(u, v)$ and variance σ^2. From (2) we have

$$\mathbb{E}[\mathcal{V}(a, b, \theta)] = \sum_{(u,v)\in\mathcal{E}_i(a,b,\theta)} \left\{ \tau(u, v) + (\lambda - 1) \left[\tau(u, v)\Phi\left(-\frac{\tau(u, v)}{\sigma}\right) - \sigma\phi\left(-\frac{\tau(u, v)}{\sigma}\right) \right] \right\} \tag{3}$$

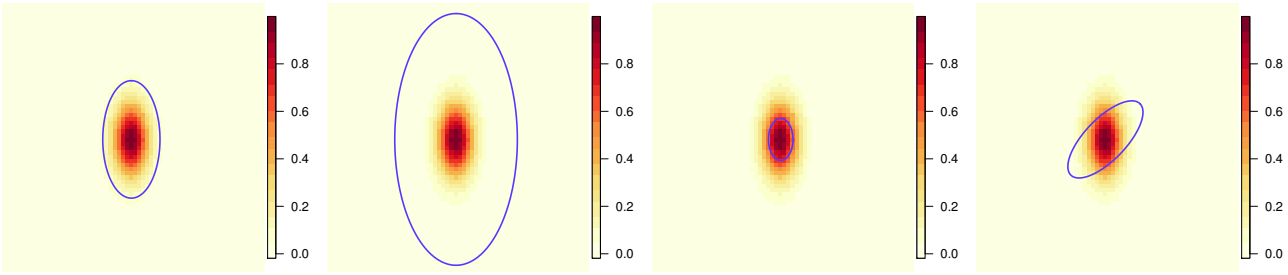
where $\Phi(\cdot)$ and $\phi(\cdot)$ are, respectively, the cumulative distribution function and the probability density function of the standard normal random variable Z. The multiplier for the term $(\lambda - 1)$ inside the summation of (3) follows from (2) since $\mathbb{E}\{Y_i^c(u, v)\mathcal{I}_{[Y_i^c(u,v)<0]}\} = \sigma\mathbb{E}[Z\mathcal{I}_{[Z<-\tau(u,v)/\sigma]}] + \tau(u, v)\Phi(-\tau(u, v)/\sigma)$ and then noting that for any $\zeta \in I\!R$, $\mathbb{E}\{Z\mathcal{I}_{[Z<\zeta]}\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\zeta} z\phi(z)dz = \phi(\zeta)$, upon substituting $w = z \exp(-z^2/2)$ in the integrand. For ease of presentation, let us use $C_\sigma(t) = t + (\lambda - 1)\left[t\Phi\left(-t/\sigma\right) - \sigma\phi\left(-t/\sigma\right)\right]$. Thus, $C_\sigma(\tau(u, v))$ is the contribution at the (u, v)th pixel to the summation in (3). Note that $C_\sigma(t)$ is a monotone increasing function in t with first derivative $C'_\sigma(t) = 1 + (\lambda - 1)\Phi(-t/\sigma)$. For any pixel outside the true signal region, $\tau(u, v) = 0$, so that its contribution $C_\sigma(\tau(u, v))$ to the summation in (3) is $-(\lambda - 1)\sigma/\sqrt{2\pi}$ which is negative as long as $\lambda > 1$. For pixels with positive $\tau(u, v)$, the sign of $C_\sigma(\tau(u, v))$ depends on the exact value that $\tau(u, v)$ takes relative to σ and λ. Let $\xi = \min\{\tau(u, v)/\sigma : \tau(u, v) > 0\}$ and let λ_ξ be the root of the equation

$$h_\xi(\lambda) = \xi\left[1 + (\lambda - 1)\Phi(-\xi)\right] - (\lambda - 1)\phi(-\xi), \tag{4}$$

which is obtained by dividing $C_\sigma(t)$ by σ and setting $\xi = t/\sigma$. Then, $\forall\lambda \in (1, \lambda_\xi)$, $C_\sigma(\tau(u, v)) > 0$ for all (u, v) in the signal region. We illustrate our ideal ellipse using a simulated image (after processing with a matched filter) in Figure 3a. Note that after allowing for the effect of pixelization, the optimal ellipse is perfectly aligned with the boundary of the signal region. Further, all terms in the summation of (3) are positive because all pixels with a negative $C_\sigma(\tau(u, v))$ are outside the drawn ellipse and so are excluded from the calculations. Also, the assumption of no more than one signal region means there are no pixels outside the ellipse with $C_\sigma(\tau(u, v)) > 0$. We now show that, under these assumptions, there are three overarching cases that we need to consider (because all other cases are subsumed within

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

these three possibilities), but all these cases result in a lower $\mathcal{V}(a, b, \theta)$ on the average.



(a) Optimal ellipse (b) Over-sized ellipse (c) Under-sized ellipse (d) Misaligned ellipse

Figure 3: (a) Sample elliptical signal region with drawn optimal ellipse matching the boundaries. (b) Over-sized, (c) under-sized and (d) misaligned ellipses that demonstrate the possibilities that justify the development of the specific $\mathcal{V}(a, b, \theta)$ of Section 2.2.1.

Case 1 - Over-sized ellipse drawn beyond the boundary of the signal region: Here, as displayed in Figure 3b, the drawn ellipse covers all of the signal region, but also contains a substantial amount of area with no signal. In this case, the positive terms that contribute to the summation in the optimal ellipse of Figure 3a get depleted by the (noise) pixels outside the signal region that are included inside the ellipse because the contribution $C_\sigma(\tau(u, v))$ of these noise pixels to $\mathcal{V}(a, b, \theta)$ is negative. Thus, in expectation, the $\mathcal{V}(a, b, \theta)$ is less than that for Figure 3a.

Case 2 - Under-sized ellipse drawn inside the signal region: Figure 3c illustrates the case where the drawn ellipse is inside the signal region but some signal pixels lie outside the ellipse and are considered as noise. Here, two factors reduce $\mathbb{E}[\mathcal{V}(a, b, \theta]$. Primarily, there are fewer positive terms $C_\sigma(\tau(u, v))$ included in the summation than for the case with Figure 3a because pixels in the signal region but outside the drawn ellipse have been excluded from the summation. Further, the estimate $\bar{Y}^i$ of μ is biased upwards because it includes pixels with $\tau(u, v) > 0$. Consequently, $\mathbb{E}[Y^i(u, v)] \leq \tau(u, v)$ and because $C_\sigma(t)$ is a monotonically increasing function, each of the included contributions to the summation in (3) is also potentially reduced. Thus, we also get a lower $\mathbb{E}[\mathcal{V}(a, b, \theta)]$ than in Figure 3a.

Case 3 - Ellipse drawn with misaligned axis: The drawn ellipse is misaligned, which means that the orientation of its principal axes is at variance with that of the (elliptically-shaped) signal. As a result, some signal pixels lie outside the ellipse, while some others that are not part of the signal region are included inside it. Thus, the terms in the summation of (3) include some $C_\sigma(\tau(u, v))$-values that are negative (these are the noise pixels with $\tau(u, v) = 0$) while also excluding some pixels for which $C_\sigma(\tau(u, v))$ has a positive contribution (signal region pixels that are outside the ellipse, resulting in the same reduction of $\mathbb{E}[\mathcal{V}(a, b, \theta)]$ as in Case 2 above). In this case also, $\mathcal{V}(a, b, \theta)$ is expected to be lower than in the case where the ellipse matches the boundary of the signal region (Figure 3d).

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

The discussion here provides some theoretical grounding for the use of $\mathcal{V}(a, b, \theta)$ as an objective function to be maximized in terms of the inner ellipse parameters (a, b, θ). We have shown that for over-sized, under-sized or disoriented ellipses, the volume criterion is, on the average, smaller than that for the case where the ellipse tracks the boundary of the signal region. Thus, maximizing $\mathcal{V}(a, b, \theta)$ as a function of (a, b, θ) is a necessary step to obtain the optimal ellipse. Finally, we restrict our search for (a, b, θ) to be in a closed set to guarantee the existence of a maximum for a given value of λ.

A reviewer has asked about the algorithm used to obtain the global maximum. Like with many other iterative optimization algorithms, convergence to a maximum depends on the initializing values. We take some care in choosing these initializers by first using a 3-D grid of starting values for (a, b, θ) in $(10 \times 10 \times 10)$-grid and then, evaluating $\mathcal{V}(a, b, \theta)$ at each grid value. The combinations providing the five largest value of (2) are each run for ten "short" iterations. The best performer at the conclusion of these ten "short" iterations is then used to start the optimization algorithm and run to convergence.

Our initialization algorithm and convergence assessmment is a hybrid adaptation of the *em-EM* (Biernacki et al., 2003) and and *Rnd-EM* (Maitra, 2009) algorithms used in the context of intializing the Expectation-Maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) for parameter estimation in Gaussian mixture models. In *em-EM*, the EM algorithm is initialized at several (random) initial values and then run to lax convergence with each initializer for a fixed total number of iterations. The solution providing the highest loglikelihood value at lax convergence is then run to strict convergence. The *Rnd-EM* algorithm trades off the lax convergence steps in *em-EM* by eliminating it in favor of a (large) total number of initializers at which the loglikelihood is evaluated, with the solution that produced the highest loglikelihood value then run to strict convergence. Both these methods have been shown to be competitive initializers (Maitra and Melnykov, 2010) with no clear winner so we adapt a hybrid scheme of both in initializing our optimization algorithm. As a check, we have reviewed contour plots of the volume function for several sample vibrothermography datasets and found no cause for concern about the existence and identification of the global maximum.

2.2.2 Effect of λ

The objective function (2) depends on the regularization parameter λ: Figure S-1 illustrates that a larger λ value leads to smaller ellipses and vice-versa. We now invoke (4) in developing guidelines for choosing λ. The root of (4) is given by $\lambda_\xi = \xi/[\phi(-\xi) - \xi\Phi(-\xi)] + 1$ so that the choice of λ_ξ depends on ξ which is the minimum intensity of the signal region relative to the noise standard deviation σ. Consequently, ξ can be viewed as the minimum value that a flaw signature can be expected to have in order to ensure a high POD. This is context-dependent and should ideally be set according to a standard determined by the particular application. Our discussion above has shown that we have $C_\sigma(\tau(u, v)) >$

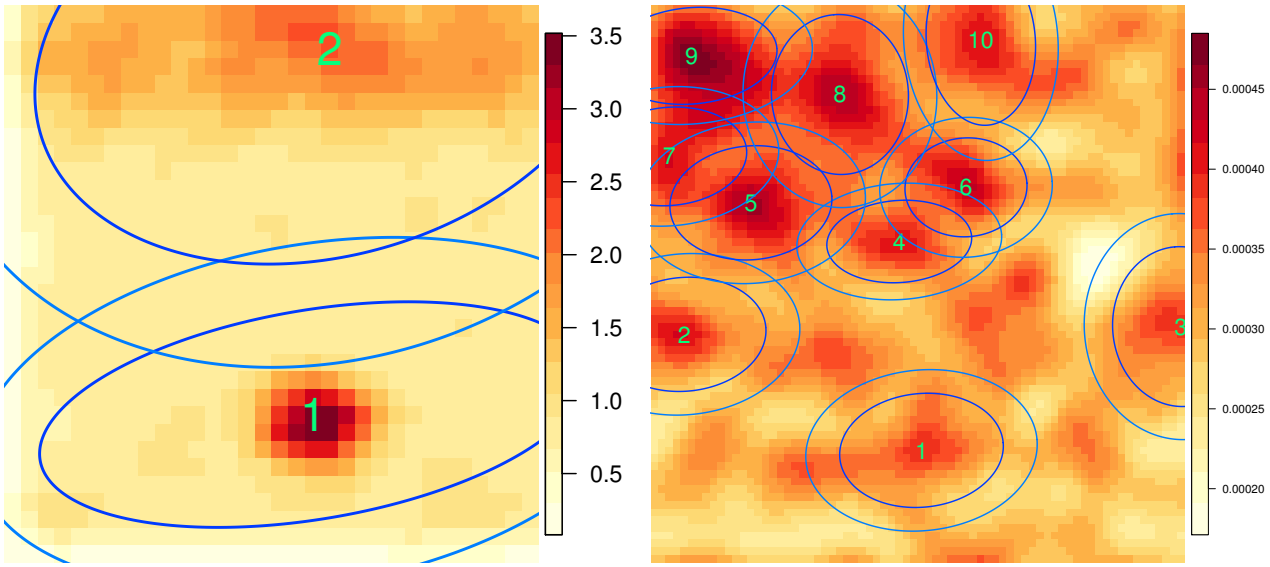
This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

0 as long as $\lambda \leq \lambda_\xi$. However, given that only flaws of intensity greater than $\xi\sigma$ are expected to be of importance in our application, we recommend setting $\lambda \equiv \lambda_\xi$. We emphasize that our recommendation of λ_ξ is only a guideline: our derivation above has invoked the distributional assumption of Gaussian white noise, which may not be accurate for the processed image. We have found, however, that the exact value of λ around a given λ_ξ does not appreciably alter the results. In particular, there is not much difference in results for λ_ξ-values obtained over a range in $(\lambda_{z_{0.95}}, \lambda_{z_{0.975}}) \equiv (79.73, 208.49)$. For our experiments reported in this paper, we have used $\lambda = 100$, simply as a matter of choice.

2.2.3 Some extensions of our algorithm

Our algorithm has so far been developed for one (elliptical) signal region. Some NDE imaging applications could have multiple hotspots while some others could have a pair of hotspots (*e.g.*, from a crack with two tips) generated from a single flaw. We now extend our algorithm for these cases.

Extension allowing for multiple flaws: Figure 4a displays an ultrasound image after processing with a matched filter to illustrate a case with potential non-negligible probability of having multiple flaws. Briefly, ultrasound images are used to find near-surface and subsurface flaws in materials by transmitting ultrasonic waves through them. Specifically, Figure 4a is an image of a synthetic inclusion forging disk (known as SID) as described in Section 6.2 of Margetan et al. (2007). This particular



(a) (b)

Figure 4: Detection of multiple potential targets in (a) an image ultrasound and (b) a vibrothermography image of a noise specimen using our proposed algorithm.

SID contains numerous types of synthetic hard alpha (SHA) inclusions and flat bottom holes (FBHs) of different known sizes. Additionally, for titanium and other similar noisy materials with large grain boundaries, there is also noise as a consequence of banding, leading to a combination of low- and

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

high-noise regions. In the illustration, the disk is known to have an SHA as well as signal from a high-noise region of the SID. Our objective is to extend our algorithm to make it possible to identify a flaw in the presence of multiple hotspots (the SHA and the banding noise in this example) and to evaluate their SNRs for analysis. We do so by incorporating the additional steps:

1. First, we detect centers for all possible indications under the assumption that each signal region is elliptical, with intensity decreasing away from the center of the region where it is highest. For each such region, draw a square of 3×3 pixels and define it as a possible flaw if the intensity at its center is not smaller than at all its surrounding 8 pixels. By looking at all such 3×3-pixels moving windows in the image, we can identify all possible indications.

   Comment: The use of the 3×3-pixels window in the manner described here is designed to filter out potential candidates that are simply noise pixels with high observed intensities. If the center of a 3×3-pixels window has the highest intensity in relation to all the other eight pixels that surround it, then it is much more likely to be the center of a true flaw.
2. For each indication detected in the previous step, quantify the peak amplitude (*i.e.*, the pixel intensity at the center). Calculate the scaled amplitude as the ratio of the amplitude to the maximum amplitudes for each such candidate indication.
3. Each of the indications obtained in the previous steps are candidate flaws whose exact status needs to be determined. A number of these candidates are noise artifacts with virtually no chance of being a true flaw. We reduce computation time (in subsequent evaluation and processing stages) by eliminating such candidates. Specifically, we eliminate all those candidates with scaled amplitude less than a pre-set threshold $\varrho \in (0, 1)$. In the experiments of this paper, we set $\varrho = 0.9$.
4. As a final step, maximize the volume (2) on each indication selected in Step 3 to draw the optimal inner and outer ellipse for feature extraction on the selected indications. Thus, we get a set of indications, each of which is a potential flaw that needs further analysis for accurate determination.

We refer back to Figures 4a (and b) to illustrate the performance of our algorithm in detecting multiple flaws. Our algorithm identified two hotspots indicated by "1" and "2" in Figure 4a in the decreasing order of their scaled amplitudes. (The inner and outer ellipses for SNR calculations are also drawn around each indication.) Indication "1" in the image corresponds to a flaw arising out of an SHA while indication "2" corresponds to the banding background titanium noise mentioned earlier. Figure 4b illustrates performance of our algorithm on a noise specimen in vibrothermography. Note that the fact that there is no true signal in an image from a noise specimen means that there is greater potential for identification of multiple indications, and indeed Figure 4b does identify multiple possible targets. (Once again, we have drawn the inner and outer ellipses around each potential indication.) We conclude

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

our discussion of this illustration by noting that although two candidate flaws were selected in Figure 4a using the steps above, only indication "1" gives a relatively high SNR of 7.90. The SNR for indication "2" is 2.03, suggesting that it is likely a noise artifact. For Figure 4b, we identified 10 possible targets but these all had SNRs below 2, indicating that it is likely that all these are noise artifacts. We provide a more formal approach to deciding on the detection limit for SNR-based metrics in Section 2.3, noting here only that the extension of our algorithm can successfully accommodate multiple flaw detection.

Extension to allow for signal in the form of a pair of hot regions: In some applications, including vibrothermography, flaws may develop and be imaged in the form of two indications that are close to each other with a bright region in between, as shown in the vibrothermography image in Figure 5a of a seeded flaw after processing with a matched filter. In this particular example, there is only one

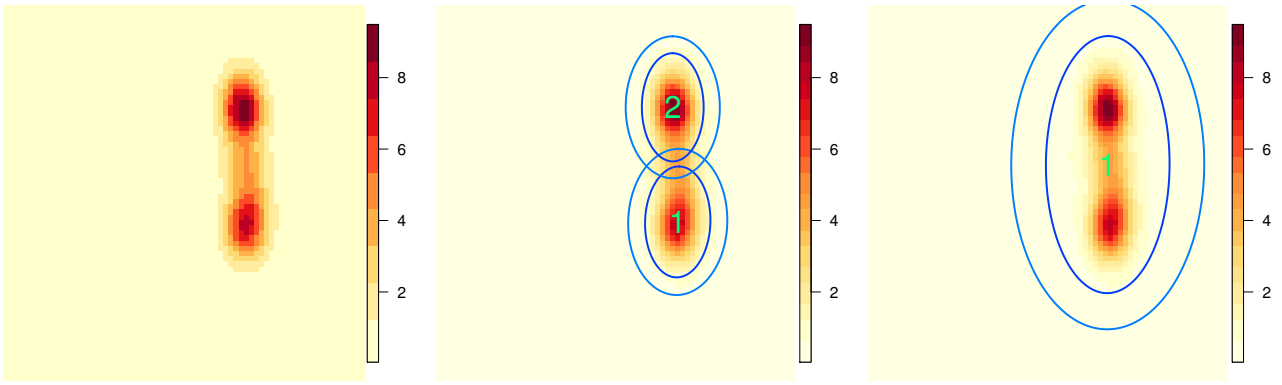


Figure 5: (a) Processed vibrothermography image with two hot tips in one indication region and illustration of (b) algorithm for detecting multiple hotspots and (c) its extension for deducing that the two hotspots are in reality from one signal souce.

crack but heat is generated from the extremal crack tips. For such cases, although we may be able to draw a set of ellipses for each of the two hotspots, the SNR's may be low. This is because the outer elliptical frames contain some high-intensity pixels, thus enhancing $\bar{e}$ and attenuating the SNR. In order to address this situation, we can consider two peaks as belonging to a single indication and create a new center as the midpoint of centers of the two original ellipses. Then we can draw a new set of ellipses covering both signal indications and extract useful metrics. The additional algorithmic steps are as follows:

1. First, for any pair of hotspot indications detected using the methodology and extension described above, check if the two centers are close to each other. The threshold to detect closeness can be pre-specified in terms of the number of pixels, corresponding to the largest expected flaw size.
2. If the two centers are close, with corresponding SNR's that are smaller than some pre-specified SNR threshold (*e.g.*, one may adopt 2.5 as the SNR threshold following the standard used for ultrasonic

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

flaw detection in titanium parts), use the midpoint of the two centers as the new center, and draw another (larger) set of ellipses that covers both indications "1" and "2" (both hotspots are now considered to be part of the same indication). If the new SNR (corresponding to the larger ellipse) is greater than both the original SNRs, we use the new region (larger ellipse) to replace the original two indication regions (smaller ellipses).

Figures 5b and c illustrate the application of the original algorithm and the above extension on the image of Figure 5a. Using the original algorithm without pairwise correction, two sets of ellipses can be nicely drawn, but these have relatively low SNR values (2.44 for indication "2" and 2.12 for indication "1"). These SNR's are below the standard threshold SNR of 2.5 so we cannot claim these indications individually as flaws, even though the existence of the flaw is visually clear. However, if we apply the extension, the algorithm correctly determines that the two centers are close, leading to a new round of optimization performed after combining the two signal regions together. The resulting SNR for the larger sets of ellipses is 36.91, large enough to claim that a flaw has been detected.

2.3 Detection rule and statistical models for estimating POD

2.3.1 Choice of a detection threshold

From the development in Section 2.2, we can extract some important metrics from every indication (defined by the inner and outer ellipses). These metrics – the signal peak $\check{Y}$, noise peak $\check{e}$ and average noise $\bar{e}$ – are representative of the image and the flaw and can be used to calculate the SNR according to the definition in Section 2.1.2. Detection of a flaw is claimed whenever SNR $> \alpha$, where α is some SNR detection criterion that depends on the NDE technology and the application. (For instance, $\alpha = 2.5$ is typically used in multi-zone ultrasonic inspection of titanium billets and forgings.) We make α adaptive to our examples and facilitate comparisons by defining α as that value which gives an observed probability of false alarm (PFA) of some pre-specified value (e.g. 3% for the vibrothermography applications and $\leq 1\%$ for the ultrasonic inspections), based on the images corresponding to specimens without any flaws. In this paper, we follow Olin and Meeker (1996)'s definition of the PFA as the probability, for a particular inspection opportunity, of a flaw determination when there is no flaw. Operationally, α is the (1 – PFA) quantile of the SNR values for the observed noise images.

Following Nieters et al. (1995), we define a noise threshold as $e_{th} = \alpha\check{e} + (1-\alpha)\bar{e}$, which is a random threshold that varies from one potential flaw to another and that tends to be lower in regions with lower amounts of background noise. The above detection criterion is then equivalent to $\check{Y} > e_{th}$ (equivalently $D = \log_{10}\check{Y} - \log_{10} e_{th} > 0$). Thus, we have an indication (*i.e.*, we claim a flaw detection) whenever a specimen produces a positive value of D. Based on this criterion, each indication or candidate flaw in an image can be classified as either a detect (actual flaw) or a non-detect (noise artifact). In

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

our evaluations on vibrothermography noise images, we used $\alpha = 2.354$, with classification results as in Figure 6.

We use statistical modeling to describe the relationship between D and flaw size. Figure 7 is a scatterplot showing the relationship between the observed metric D and the logarithm (using base 10, as is common in many engineering applications, including NDE) of flaw size. The detection criterion metrics (D) for noise regions are also plotted in the figure and the horizontal dashed line represents the detection threshold, above which one would claim a flaw detection (a threshold of zero is used here).

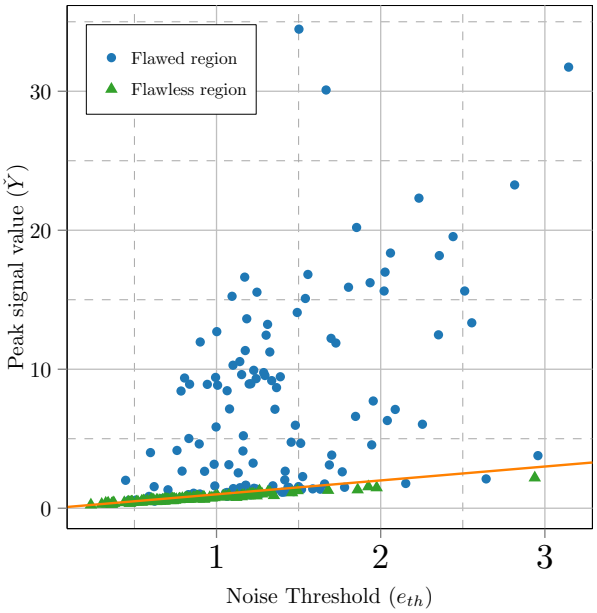


Figure 6: Classification of vibrothermography images, with detection claimed if $\check{Y} > e_{th}$, that is, if the data point is above the dashed line. Dots and triangles correspond to flawed and flawless regions respectively.

2.3.2 The Noise-Interference Model

Figure 7 shows an approximately linear relationship between D and $\log_{10}$(crack size) for specimens with $\log_{10}$(crack size) larger than 1.56, (*i.e.*, signal data points that are not in the highlighted rectangular region). This linear relationship levels off as $\log_{10}$(crack size) falls below 1.55, corresponding to the D-values in the highlighted rectangle. These D-metrics (in the highlighted rectangular region) tend to have the same level as those for noise specimens (represented by the ▲ symbols). This leveling-off behavior can be explained by the fact that small flaws may be swamped by surrounding noise, so we conclude that the observed response has a high probability of having been caused by a noise artifact even in the presence of a small flaw. Li and Meeker (2009) introduced the noise interference model (NIM) to describe this kind of relationship between a response and flaw size. According to this

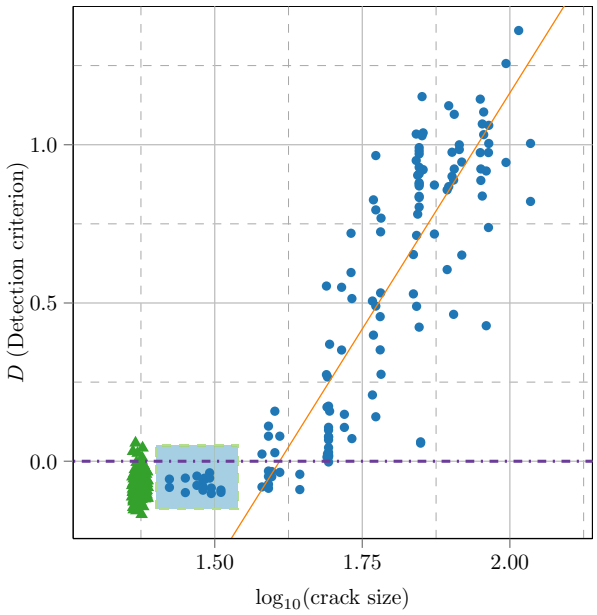


Figure 7: Plot of D (obtained after drawing the optimal ellipses against log_{10}(flaw size) along with the fitted regression line based on the NIM model.

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

model for the vibothermography example (where the flaws are cracks), we consider the observed D-metric corresponding to a specimen with signal to be the competing result of signal (D_{signal}) and noise (D_{noise}), *i.e.*, $D_{\text{obs}} = \max(D_{\text{signal}}, D_{\text{noise}})$ where $D_{\text{signal}} = \beta_0 + \beta_1 \log_{10}(\text{crack size}) + \varepsilon_s$, where ε_S follows a normal distribution $N(0,\ \sigma_S^2)$ and $D_{\text{noise}} \sim N(\mu_N,\ \sigma_N^2)$. Based on this probabilistic model specification, the observed likelihood for specimens with a crack is given by

$$L_{\text{crack}} = \prod_{\text{k with crack}} \sigma_S^{-1}\phi\left[(D_k - \beta_0 - \beta_1 \times \log_{10}(\text{crack size}))/\sigma_S\right] \Phi\left[(D_k - \mu_N)/\sigma_N\right] + \sigma_N^{-1}\Phi\left[\sigma_S^{-1}(D_k - \beta_0 - \beta_1 \log_{10}(\text{crack size}))\right] \phi\left[\sigma_N^{-1}(D_k - \mu_N)\right] \tag{5}$$

while the likelihood for flawless specimens is $L_{\text{noise}} = \prod_{\text{j without flaw}} \sigma_N^{-1}\phi\left[\sigma_N^{-1}(D_j - \mu_N)\right]$. Because specimens are independently inspected, the total likelihood for all the specimens is, as usual, the product of the two likelihoods $L = L_{\text{crack}} L_{\text{noise}}$, from which maximum likelihood estimates (MLEs) of $\beta_0, \beta_1, \mu_N, \sigma_S$, and σ_N can be obtained in the usual way. The resulting fitted regression line for the NIM model after plugging in the MLEs for our vibrothermography datasets is shown in Figure 7. Finally, the normality assumption in the NIM seems reasonable (see Section S-2.2 for details).

3 Performance Evaluations

3.1 Simulation Experiments

We evaluate performance of our algorithm for automated optimal feature extraction as well as the use of the NIM on the D-metrics obtained from the extracted features. Performance calibrations are in terms of the capability of our methods and modeling to detect flaws of different sizes.

3.1.1 Performance Metrics

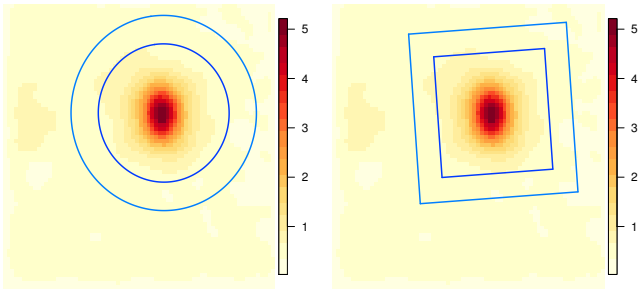
We follow the NDE practice of quantifying detection capability in terms of the POD which is also the most commonly-used metric in the NDE and statistical literature. According to the NIM model of Section 2.3.2, we can (see the derivations in Section S-3.1) express the POD as a function of flaw size and the regression parameters using the model:

$$\text{POD(flaw)} = \Pr(D_{obs} > 0) = 1 - \Phi\left[-\frac{\beta_0 + \beta_1 \log_{10}(\text{flaw})}{\sigma_S}\right] \Phi\left(-\frac{\mu_N}{\sigma_N}\right). \tag{6}$$

This definition of POD takes credit for a detection for cases where a noise artifact results in a signal stronger than that from the actual flaw and results in an above-threshold observed signal. Our performance evaluations compare our proposed algorithm with alternatives in terms of the POD. These alternatives employed (a) the commonly-used "peak amplitude" method described in Section 1.2 and

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

(b) an automated and optimized version of the Howard and Gilmore (1994) rectangle-drawing method which uses a development parallel to Section 2.2.1 except that it is used to draw an optimal inner rectangle. We call this method "Rectangle" and note that it is expected to be as good, if not better, than an operator-drawn rectangular set (which is subjective and does not involve optimality considerations). For easy reference, we use "Ellipse" to refer to our proposed method based on drawing optimal inner and outer ellipses. Figure 8 displays a comparison of the results obtained using inner and outer ellipses (left) and inner and outer rectangles (right). Similar to the developments in Section 2.2.1, the "Rectangle" method is also readily extended for multiple targets and single targets with paired hotspots. The POD impression for the "Rectangle" method is the same as that for our "Ellipse" method and is given by (6). The expression is, however, not applicable to the "peak amplitude" method (henceforth abbreviated as "PeakAmp") so we develop the POD for this case next.



(a) (b)

Figure 8: Results of using the algorithm on feature extraction using the proposed algorithm with (a) drawn optimal inner and outer ellipses and (b) drawn optimal optimal inner and outer rectangles.

The "PeakAmp" method uses the peak intensity $\check{Z}$ *of the raw unprocessed hottest frame image* to characterize the image. Here we have a signal pixel Y_S related to the flaw size by the model $\log_{10} Y_S = \gamma_0 + \gamma_1 \log_{10}(\text{flaw}) + \upsilon_S$, where the parameter γ_0 controls the magnitude of the hottest pixel, γ_1 reflects the relationship between peak amplitude and flaw size, and $\upsilon_S \sim N(0, \kappa_S^2)$. For a noise pixel, we have $\log_{10} Y_N \sim N(\nu_N, \kappa_N^2)$. This relationship was inspired by the finding in many of our (real-life) datasets that the logarithm of peak amplitude and the logarithm of flaw size is approximately linearly-related. For some inspection methods, including ultrasound, there is a physics-based explanation for this relationship, as described in Li et al. (2014). Using the NIM model to describe the relationship between the model response (*i.e.*, $\log_{10} \check{Z}$) and flaw size, the POD (for detailed derivation, see Section S-3.1 or Equation (3) of Li and Meeker, 2009) is

$$\text{POD}\,(\text{flaw}) = \Pr\left(\check{Z} > Z_{\text{th}}\right) = 1 - \Phi\left(\frac{\log_{10}(Z_{\text{th}}) - \gamma_0 - \gamma_1 \log_{10}(\text{flaw})}{\kappa_S}\right) \Phi\left(\frac{\log_{10}(Z_{\text{th}}) - \nu_N}{\kappa_N}\right),$$

where Z_{th} is the detection threshold for the peak-amplitude response and $(\gamma_0, \gamma_1, \kappa_S, \nu_N, \kappa_N)$ are estimated using the MLEs in the same manner as before.

3.1.2 Results

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

Figure 9 compares the POD curves corresponding to the "Ellipse" algorithm with those obtained using the "Rectangle" and "PeakAmp" methods. The PFA-values for all methods have been calibrated to be the same (0.03 here) to allow a fair comparison. The flaw size at which each curve attains a POD of 0.9 (indicated by the horizontal dashed line) is called the a_{90} value in NDE parlance. In the NDE community, the estimated a_{90} is widely used as a scalar metric of inspection capability and to compare different NDE methods. A smaller a_{90} value is usually a sign of a better, more sensitive, procedure. Figure 9 thus shows that both the SNR-based "Ellipse" and "Rectangle" algorithms outperform the simpler "PeakAmp" method by yielding better POD curves and smaller a_{90} values. Further, "Ellipse" has somewhat higher POD values than "Rectangle" for almost the entire range of the flaw sizes in consideration. The a_{90} estimates for the "Ellipse", "Rectangle" and "PeakAmp" methods are 52.2 mils, 53.8 mils and 61.0 mils, respectively. These a_{90} values suggest that the proposed "Ellipse" method performs better than the two other methods, indicating significant improvement over the alternatives. Note that the most important part of the POD curve, from a practical perspective, is where probability of detection is relatively high. A common saying in the NDE field is "It is not the smallest flaw we might detect that is of interest, but rather the largest flaw we might miss." Under this adage, "Ellipse" is a better performer than the "Rectangle" or "PeakAmp" methods.

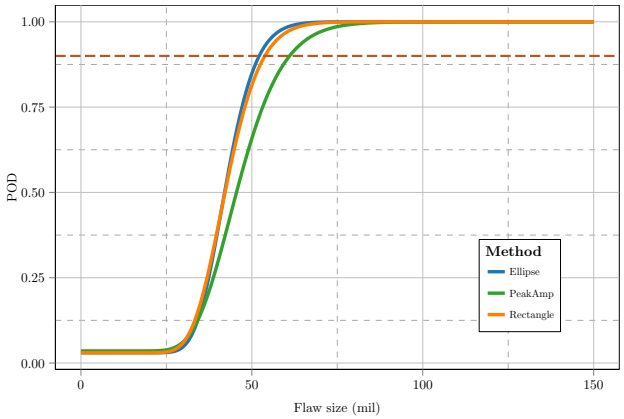


Figure 9: POD curves for the proposed algorithm and the two alternative methods.

We also evaluated the performance of our algorithm beyond the original datasets on a series of simulated NDE images. The simulated images are composed of two parts: the signal and the noise. The noise images for our simulations were obtained by resampling (with replacement) from the pool of flawless vibrothermography images (also called noise images). For our simulated flawed images, we added – to these resampled noise images – signal images from a Gaussian signature with peak amplitude defined as before, that is, $\log_{10}(\text{peak amplitude}) = \gamma_0 + \gamma_1 \log_{10}(\text{flaw})$. The FWHM $\hbar$ of the Gaussian signature is proportional to the flaw size where the constant of this proportionality is given by k. In our simulations, we used $\gamma_0 = -7$, $\gamma_1 = 1.22$ and $k = 0.0785$. For each of the flaw sizes for the specimens used in the vibrothermography inspection, we obtained 20 simulated degraded signal images and applied the three detection methods to them as well as to the noise images. Thus, three POD curves were produced and corresponding a_{90} values were computed to compare the three methods. The resulting 20 sets of POD curves are shown in Figure 10. Clearly, the "Ellipse" method

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

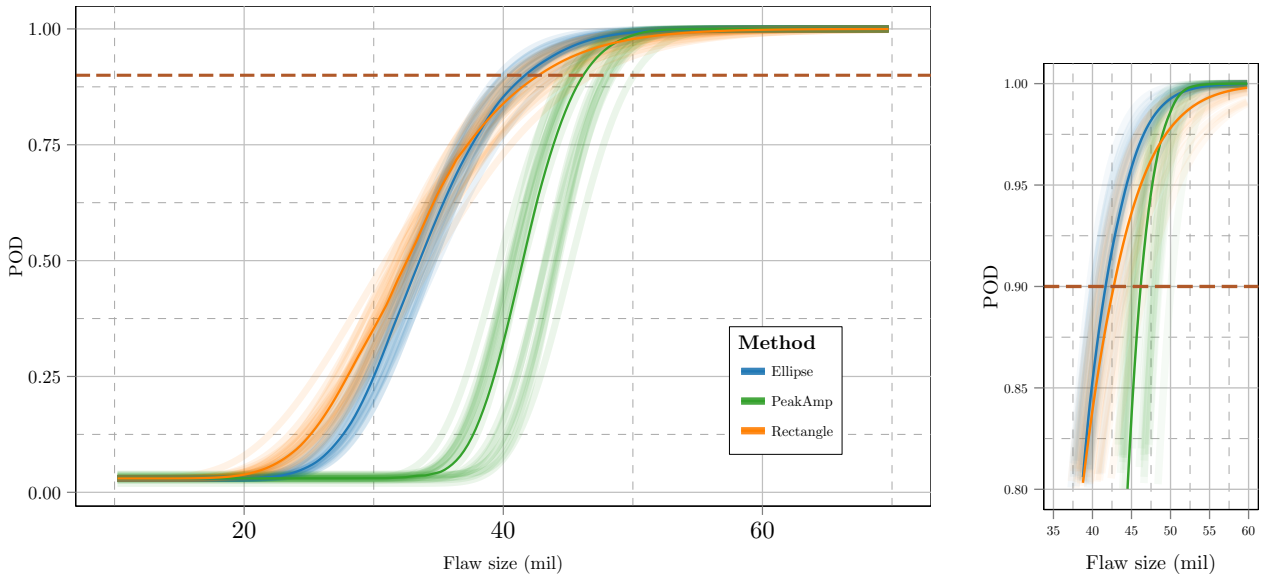


Figure 10: (a) POD comparison of three methods based on 20 simulations and (b) a closer view of the POD comparison for flaw sizes between 35 and 60 mils.

generally produces better POD curves than the "Rectangle" method but both approaches produce better POD curves than the "PeakAmp" methods.

Figure 11 displays estimated a_{90} values obtained from the three methods. We see relatively good separation among the three a_{90} distributions. Also, "Ellipse" tends to produce the smallest a_{90} values. Thus, in general, it performs better than "Rectangle" which in turn performs better than "PeakAmp". The parallel coordinates plot underlying the boxplots indicates better performance of "Ellipse" over "Rectangle" for all but two cases where "Rectangle" performs slightly better. Both of these methods are almost always better than "PeakAmp" (with the "Rectangle" method being slightly outperformed in only one case, and "Ellipse" being always superior). This assertion was confirmed by the results of a paired Wilcoxon signed rank test between the a_{90}-values

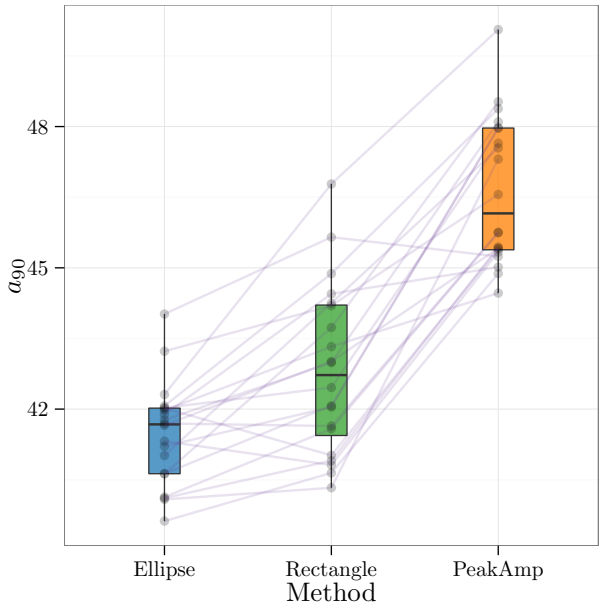


Figure 11: Estimates of a_{90} obtained by the three methods. The parallel coordinates plot connects the estimates obtained for the three methods from each experiment.

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

obtained using "Ellipse" and "Rectangle" (which produced a p-value of 0.0002 and a similar test between the a_{90}-values obtained using "Rectangle" and "PeakAmp" a_{90}-values (p-value=3.82×10^{-6}). These simulation results are consistent with what we observed in the real vibrothermography datasets. Taking into consideration both the real vibrothermography data analysis and the simulation results, we are confident that the proposed "Ellipse" method is a strong candidate for analyzing NDE image data, in the sense of producing high POD. We now analyze the dataset introduced in Section 1.1.

3.2 Application to Vibrothermography Samples

The three raw images introduced in Figure 1 with postprocessed (matched filtered) versions displayed in Figure 2 were analyzed using the methods illustrated in this paper. Under the "Ellipse" method, the SNR for the image with the strong signal (Figure 1a) was computed to be 16.12, while that for the weak signal (Figure 1b) was 5.56 and the noise image (Figure 1c) of no true flaw was calculated to be 2.30. No matter whether we use the threshold of $\alpha = 2.354$ as obtained using the development in Section 2.3.1 above, or the industry standard of 2.5, the first two images are correctly classified as signal images (which means that they contain flaws, which were cracks in this application) and the last image is correctly classified as solely containing noise (or image of a flawless specimen). Under the "Rectangle" method, the SNRs were computed to be 14.52, 5.77 and 2.37 for the three images in Figure 1a, 2b and 1c, respectively. Using developments similar to Section 2.3.1, but for the "Rectangle" method, we got the cut-off of 2.5 (which is the same as the industry standard threshold) and therefore the conclusions for these three specimens match the ones for the "Ellipse" method. We stress here that the automated and optimized version that is our "Rectangle" method, and which by itself is a contribution of this paper, is not actually used in NDE; current NDE practice utilizes subjective operator-assisted rectangle-drawing. Using the "PeakAmp" method however, the values obtained (of the pixel with the highest raw intensity) are 0.35, 0.09 and 0.053 for the three images in Figures 1a, b, and c. The cut-off value for the "PeakAmp" method to separate flaw from noise was derived using the methods in Section 2.3.1 as 0.104, so only the image with the strong signal is identified as such and the method is unable to distinguish between a weaker signal and noise, classifying both as noise specimens. Our results provide confidence in the applicability of the "Ellipse" (and "Rectangle") methodology for automated flaw detection in vibrothermography images. We note also that the SNR for the first image is much larger than that for the second image, a finding that is in good agreement with the fact that the first specimen has a larger crack than the second specimen.

4 Discussion

In this paper we developed an automated flaw detection algorithm based on image processing and SNR detection. By setting the same probability of false alarm (PFA), the proposed automated

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

algorithm based on drawing (optimal) ellipses around the signal shows better detection performance than the alternative that is based on a simplistic peak-amplitude scalar response or an intermediate algorithm that is an automated and optimized version of a method where the operator draws rectangles based on visual inspection. The POD values tend to be higher for the proposed algorithm than the other two methods for the flaw size range of interest. Correspondingly, the a_{90} value for the proposed algorithm is smaller than that of the other two algorithms. The simulation results based on simulating Gaussian-like signal and resampling of the vibrothermography noise images confirm that the proposed algorithm outperforms the other algorithms.

There are a number of areas that would benefit from increased attention. Our methods here followed the industry standard in adopting the hottest frame for analysis. It would be interesting to see if some other approach can provide similar or better results. We have also here demonstrated and evaluated performance in terms of the analysis of vibrothermography (and while not presented in detail in this paper, ultrasound) images. It would also be important to see if our methodology can be adapted and extended to other NDE imaging inspection methods such as eddy current or X-ray inspection. We expect that the answer is in the affirmative, but this will need to be validated and tested. Finally, we hope that our results will also spur greater statistical interest and involvement in this important industrial field of application.

Acknowledgements

This work was performed with partial support from the Federal Aviation Administration under contract number DTFACT-09-C-00006 through the Center for Nondestructive Evaluation at Iowa State University. The ultrasonic test images used in Section 2.2.3 were acquired as part of a project to study the POD of ultrasonic inspections of forgings, supported by the Federal Aviation Administration under contract number 08-C-00005 to Iowa State University. We thank Tim Gray for providing the images and for helping us to interpret them. We also thank the Editor, an Associate Editor and reviewers for suggestions that helped us to improve an earlier version of this paper.

References

Annis, C. and Erland, K. (1989), "Measuring differences among probability of detection curves," in *Review of Progress in Quantitative Nondestructive Evaluation*, eds. Thompson, D. O. and Chimenti, D. E., New York: Plenum Press, vol. 8, pp. 2229–2234.

Aoki, K. and Suga, Y. (1999), "Application of artificial neural network to discrimination of defect type automatic radiographic testing of welds," *ISI International*, 39, 1081–1087.

Beck, J. V., Cole, K. D., Haji-Sheikh, A., and Litkouhi, B. (1992), *Heat Conduction Using Green's Functions*, Taylor and Francis.

Berens, A. P. and Hovey, P. W. (1981), "Flaw detection reliability criteria," Tech. Rep. AFWAL-TR-81-4160,

Wright-Patterson Air Force Base, Ohio.

— (1982), "Characterization of NDE Reliability," in *Review of Progress in Quantitative Nondestructive Evaluation*, eds. Thompson, D. O. and Chimenti, D. E., New York: Plenum Press, vol. 1, pp. 579–585.

— (1983), "Statistical methods for estimating crack detection probabilities," in *Probabilistic Fracture Mechanics and Fatigue Methods: Applications for Structural Design and Maintenance*, eds. Bloom, J. M. and Ekvall, J. C., ASTM STP 798, pp. 79–94.

— (1984), "Flaw detection reliability criteria," Tech. Rep. AFWAL-TR-84-4022, Wright-Patterson Air Force Base, Ohio.

Bergholm, F. (1987), "Edge focusing," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 9, 726–741.

Biernacki, C., Celeux, G., and Govaert, G. (2003), "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models," *Computational Statistics and Data Analysis*, 413, 561–575.

Bray, D. E. and McBride, D. (eds.) (1992), *Nondestructive Testing Techniques*, New York: John Wiley.

Bray, D. E. and Stanley, R. K. (1996), *Nondestructive Evaluation: A Tool in Design, Manufacturing and Service*, CRC Press.

Burkel, R. H., Sturges, D. J., Tucker, W. T., and Gilmore, R. (1996), "Probability of Detection for Applied Ultrasonic Inspectors," in *Review of Progress in Quantitative Nondestructive Evaluation*, eds. Thompson, D. O. and Chimenti, D. E., New York: Plenum Press, vol. 15B, pp. 1991–1998.

Chen, C. and Wang, X. (2004), "Speckle reduction and edge enhancement of NDE C-scan images using ICA," *Review of Quantitative Nondestructive Evaluation*, 23, 573–580.

Collins, R., Michael, D. H., Mirshekar-Syahkal, D., and Pinsent, H. G. (1985), "Surface electromagnetic fields around surface flaws in metals," *Journal of Nondestructive Evaluation*, 5, 81–93.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood for incomplete data via the EM algorithm (with discussion)," *Jounal of the Royal Statistical Society, Series B*, 39, 1–38.

Dogandzic, A. and Zhang, B. (2007), "Bayesian NDE Defect Signal Analysis," *Signal Processing, IEEE Transactions on*, 55, 372–378.

Engelberg, S. (2007), *Random Signals and Noise*, New York: Taylor and Francis.

Gao, C. and Meeker, W. Q. (2012), "A Statistical Method for Crack Detection from Vibrothermography Inspection Data," *Quality Technology and Quantitative Management*, 9, 58–77.

Gauch, J. M. and Pizer, S. M. . (1993), "Multiresolution Analysis of Ridges and Valleys in Grey-Scale Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 635–646.

Gonzalez, R. C. and Woods, R. E. (2008), *Digital Image Processing*, Upper Saddle River, NJ: Prentice Hall, 3rd ed.

Gray, T. A. and Thompson, R. B. (1986), "Use of models to predict ultrasonic NDE reliability," in *Review of Progress in Quantitative Nondestructive Evaluation*, eds. Thompson, D. O. and Chimenti, D. E., New York: Plenum Press, vol. 5, pp. 911–918.

Halmshaw, R. (1982), *Industrial Radiography: Theory and Practice*, Englewood, NJ: Applied Science.

— (1991), *Non-Destructive Testing*, London: Edward Arnold, 2nd ed.

Hasanzadeh, R. P. R., Moghaddamjoo, A. R., Sadeghi, S. H. H., Rezaie, A. H., and Ahmadi, M. (2008), "Optimal

signal-adaptive maximum likelihood filter for enhancement of defects in eddy current C-scan images," *NDT & E International*, 41, 371–381.

Heller, C. (2012), *Handbook of Nondestructive Evaluation*, McGraw-Hill Professional, 2nd ed.

Henneke, E. G. and Jones, T. S. (1979), "Detection of damage in composite materials by vibrothermography," *ASTM Special Technical Publication*, 696, 83–95.

Holland, S. D. (2007), "First measurements from a new broadband vibrothermography measurement system," in *Progress in Quantitative Nondestructive Evaluation*, eds. Thompson, D. O. and Chimenti, D. E., American Institute of Physics Conference Series, vol. 894, pp. 478–483.

— (2011), "Thermographic signal reconstruction for vibrothermography," *Infrared Physics and Technology*, 54, 503–511.

Hovey, P. W. and Berens, A. P. (1988), "Statistical evaluation of NDE Reliability in the aerospace industry," in *Review of Progress in Quantitative Nondestructive Evaluation*, eds. Thompson, D. O. and Chimenti, D. E., New York: Plenum Press, vol. 7, pp. 1761–1768.

Howard, P. J., Copley, D. C., and Gilmore, R. S. (1998), "The application of a dynamic threshold to C-scan images with variable noise," in *Review of Progress in Quantitative Nondestructive Evaluation*, eds. Thompson, D. O. and Chimenti, D. E., vol. 17, pp. 2013–2019.

Howard, P. J. and Gilmore, R. S. (1994), "Ultrasonic C-Scan imaging for hard alpha flaw detection and characterization," in *Review of Progress in Quantitative Nondestructive Evaluation*, eds. Thompson, D. O. and Chimenti, D. E., vol. 13, pp. 763–770.

Jain, A. K. (1989), *Fundamentals of Digital Image Processing*, Prentice Hall.

Jansohn, R. and Schickert, M. (1998), "Objective Interpretation of Ultrasonic Concrete Image," in *7th European Conference on Non-Destructive Testing*.

Kahn, A. H., Spal, R., and Feldman, A. (1977), "Eddy-current losses due to a surface crack in conducting material," *Journal of Applied Physics*, 48, 4454–4459.

Krautkramer, J. and Krautkramer, H. (1990), *Ultrasonic Testing of Materials*, New York: Springer-Verlag, 3rd ed.

Legendre, S., Goyette, J., and Massicotte, D. (2001), "Ultrasonic NDE of composite material structures using wavelet coefficients," *NDT&E International*, 34, 31–37.

Li, M., Holland, S. D., and Meeker, W. Q. (2010), "Statistical methods for automatic crack detection based on vibrothermography sequence-of-images data," *Applied Stochastic Models in Business and Industry*, 26, 481–495.

— (2011), "Quantitative Multi-Inspection-Site Comparison of Probability of Detection for Vibrothermography Nondestructive Evaluation Data," *Journal of Nondestructive Evaluation*, 30, 172–178.

Li, M. and Meeker, W. Q. (2009), "A Noise Interference Model for Estimating Probability of Detection for Nondestructive Evaluations," *Review of Quantitative Nondestructive Evaluation*, 28, 1769–1776.

Li, M., Meeker, W. Q., and Thompson, R. B. (2014), "Physical Model-Assisted Probability of Detection of Flaws in Titanium Forgings using Ultrasonic Nondestructive Evaluation," *Technometrics*, 56, 78–91.

Lindgren, A., Shull, P. J., Joseph, K., and Hagemaier, D. (2002), "Magnetic Particle," in *Nondestructive Evaluation: Theory, Techniques, and Applications*, ed. Shull, P. J., CRC Press, chap. 4, pp. 193–260.

Maitra, R. (2009), "Initializing Partition-Optimization Algorithms," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6, 144–157.

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

Maitra, R. and Melnykov, V. (2010), "Simulating data to study performance of finite mixture modeling and clustering algorithms," *Journal of Computational and Graphical Statistics*, 19, 354–376.

Maldague, X. P. V. (2001), *Theory and Practice of Infrared Technology for Nondestructive Testing*, Wiley.

Margetan, F. J., Umbach, J., Roberts, R., Friedl, J., Degtyar, A., Keller, M., Hassan, W., Brasche, L., Klassen, A., Wasan, H., and Kinney, A. (2007), "Inspection Developments for Titanium Forgings," Tech. Rep. DOT/FAA/AR-05/46, Air Traffic Organization Operations Planning Office of Aviation Research and Development, Washington, DC.

Marr, D. and Hildreth, E. (1980), "Theory of edge detection," *Proceedings of the Royal Society*, 187–217.

Martz, H. E., Logan, C. M., and Shull, P. J. (2002), "Radiology," in *Nondestructive Evaluation: Theory, Techniques, and Applications*, ed. Shull, P. J., CRC Press, chap. 7, pp. 447–596.

McLachlan, G. and Krishnan, T. (2008), *The EM Algorithm and Extensions*, New York: Wiley, 2nd ed.

Meyer, A. W. and Candy, J. V. (2002), "Iterative processing of ultrasonic measurements to characterize flaws in critical optical components," *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 49, 1124–1138.

MIL-HDBK-1823A (2009), *Nondestructive Evaluation System Reliability Assessment*, Building 4D, 700 Roberts Avenue, Philadelphia, PA: Standardization Order Desk.

Neal, S. P. and Speckman, P. L. (1993), "Flaw Signature Estimation in Ultrasonic Nondestructive Evaluation Using the Wiener Filter with Limited Prior Information," *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 40, 347–353.

Ng, S. C., Ismail, N., Ali, A., Sahari, B., and Yusof, J. M. (2013), "Ultrasonic NDE for Internal Defect Detection in Multi-layered Composite Materials by Multi-resolution Signal Decomposition," *Journal of Applied Sciences*, 13, 87–94.

Nieters, E. J., Gilmore, R. S., Trzaskos, R. C., Young, J. D., Copley, D. C., Howard, P. J., Keller, M. E., and Leach, W. J. (1995), "A multizone technique for billet inspection," in *Review of Progress in Quantitative Nondestructive Evaluation*, vol. 14, pp. 2137–2144.

Olin, B. D. and Meeker, W. Q. (1996), "Applications of Statistical Methods to Nondestructive Evaluation (with discussion)," *Technometrics*, 38, 95–130.

O'Sullivan, F. and Qian, M. (1994), "A regularized contrast statistic for object boundary estimation - implementation and statistical evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 561–570.

Perdijon, J. (1988a), "Statistics Applied to Measurements in Ultrasonic Testing," *Materials Evaluation*, 46, 1317–1323.

— (1988b), "Statistics Applied to the Inference of Flaws in Ultrasonic Testing," *Materials Evaluation*, 46, 1666–1671.

— (1989), "Statistics Applied to the Acceptance Decision in Ultrasonic Testing," *Materials Evaluation*, 47, 812–821.

Prosser, W. H. (2002), "Acoustic Emission," in *Nondestructive Evaluation: Theory, Techniques, and Applications*, ed. Shull, P. J., CRC Press, chap. 6, pp. 369–446.

Qiu, P. (2005), *Image Processing and Jump Regression Analysis*, New York: John Wiley and Sons.

Qiu, P. and Sun, J. (2007), "Local smoothing image segmentation for spotted microarray images," *Journal of the American Statistical Association*, 102, 1129–1144.

— (2009), "Using conventional edge detectors and post-smoothing for segmentation of spotted microarray images," *Journal of Computational and Graphical Statistics*, 18, 147–164.

Reifsnider, K. L., Henneke, E. G., and Stinchcomb, W. W. (1980), *Mechanics of Nondestructive Testing: Conference on the Mechanics of Nondestructive Testing*, Plenum Press.

Rosenfeld, A. (1984), *Multiresolution Image Processing and Analysis*, Springer-Verlag.

Rosenfeld, A. and Kak, A. C. (1982), *Digital picture Processing*, vol. 2, New York: Academic Press, 2nd ed.

Rummel, W. D. (1983), "Considerations for Quantitative NDE and NDE Reliability Improvement," in *Review of Progress in Quantitative Nondestructive Evaluation*, eds. Thompson, D. O. and Chimenti, D. E., New York: Plenum Press, vol. 2A, pp. 19–35.

Shull, P. J. (ed.) (2002), *Nondestructive Evaluation: Theory, Techniques, and Applications*, CRC Press.

Silk, M. G., Stoneham, A. M., and Temple, J. A. G. (1987), *The Reliability of Non-destructive Inspection*, Bristol, United Kingdom: Adam Hilger.

Spencer, F. W. and Schurman, D. L. (1995), *Reliability Assessment at Airline Inspection Facilities, III: Results of an Eddy Current Inspection Reliability Experiment*, DOT/FAA/CT-92/12, III, Atlantic City, NJ: FAA Technical Center.

Spicer, J. M. and Osiander, R. (2002), "Active Thermography," in *Nondestructive Evaluation: Theory, Techniques, and Applications*, ed. Shull, P. J., CRC Press, chap. 8, pp. 597–643.

Sweeting, T. J. (1995), "Statistical Models for Nondestructive Evaluation," *International Statistical Review*, 63, 199–214.

Turin, G. L. (1960), "An Introduction to Matched Filters," *IEEE Transactions on Information Theory*, 6, 311–329.

— (1976), "An Introduction to Digital Matched Filters," in *Proceedings of the IEEE*, vol. 64, pp. 1092–1112.

Wang, G. and Liao, T. W. (2002), "Automatic identification of different types of welding defects in radiographic images," *NDT & E International*, 35, 519–528.

Yang, G., Tamburrino, A., Udpa, L., Udpa, S. S., Zeng, Z., Deng, Y., and Que, P. (2010), "Pulsed eddy-current based giant magnetoresistive system for the inspection of aircraft structures," *IEEE Transactions on Magnetics*, 46, 910–917.

Zaki, F. W., Abd Elnaby, M. M., Elshafiey, I. M., and Ashour, A. S. (2001), "DCT and DWT feature extraction and ANN classification based technique for non-destructive testing of materials," in *Proceedings of the Eighteenth National Radio Science Conference*, vol. 1, pp. 35–44.

Zavaljevski, N., Bakhtiari, S., Miron, A., Kupperman, D. S., Wei, T. Y. C., and Marchertas, P. (2005), "Automated Algorithms for Eddy current array probes for steam generator inspection," in *Review of Progress in Quantitative Nondestructive Evaluation*, eds. Thompson, D. O. and Chimenti, D. E., New York: Plenum Press, vol. 24, pp. 728–735.

Supplement to "A Statistical Framework for Improved Automatic Flaw Detection in Nondestructive Evaluation Images"

Ye Tian
Department of Statistics and Statistical Laboratory
Iowa State University
Ames, IA 50011
(tianye1984@gmail.com)

Ranjan Maitra
Department of Statistics and Statistical Laboratory
Iowa State University
Ames, IA 50011
(maitra@iastate.edu)

William Q. Meeker
Department of Statistics and the Center for Nondestructive Evaluation
Iowa State University
Ames, IA 50011
(wqmeeker@iastate.edu)

Stephen D. Holland
Department of Aerospace Engineering and Center for Nondestructive Evaluation
Iowa State University
Ames, IA 50011
(sdh4@iastate.edu)

S-1 Notations used in Supplement

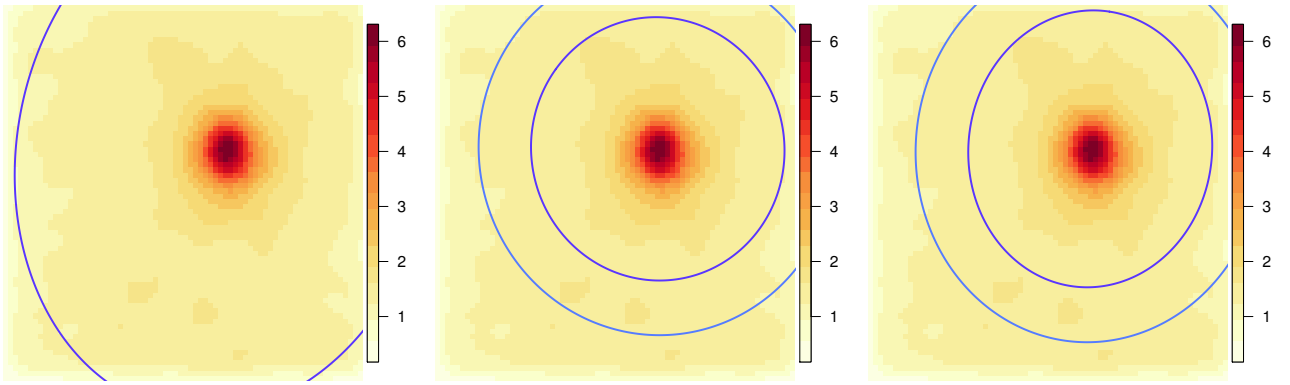
In this supplement, references to sections, figures and equations in the main paper are referred to using the same identifiers as in the main paper. References to sections, figures and equations in the supplement use the suffix "S-".

S-2 Methodology – Supplement

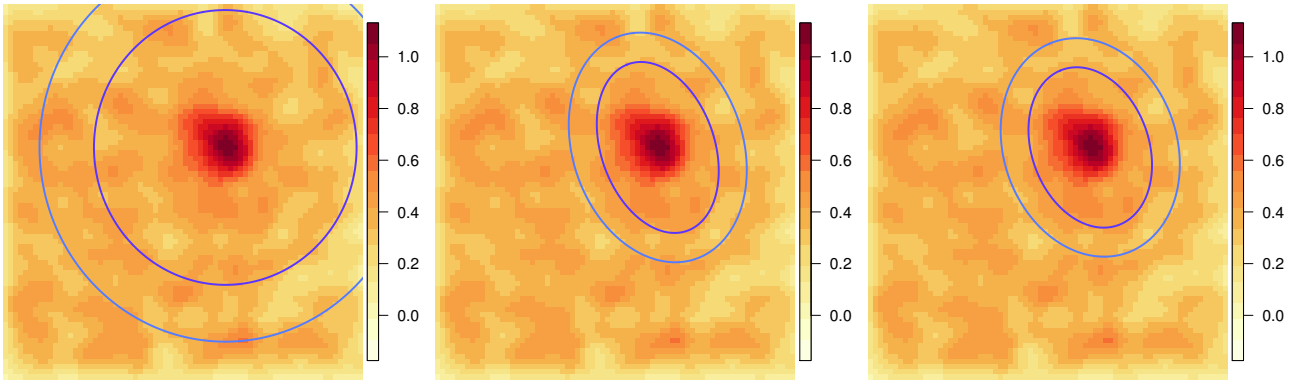
S-2.1 Illustrative Examples showing the Effect of λ

The objective function (2) in the main paper depends on λ so we now illustrate the effect of λ with a bid to make recommendations for its selection. Figure S-1 displays the results – for three different values of λ ($\lambda = 2$, first column; $\lambda = 100$, second column; $\lambda = 200$, third column) – of drawing

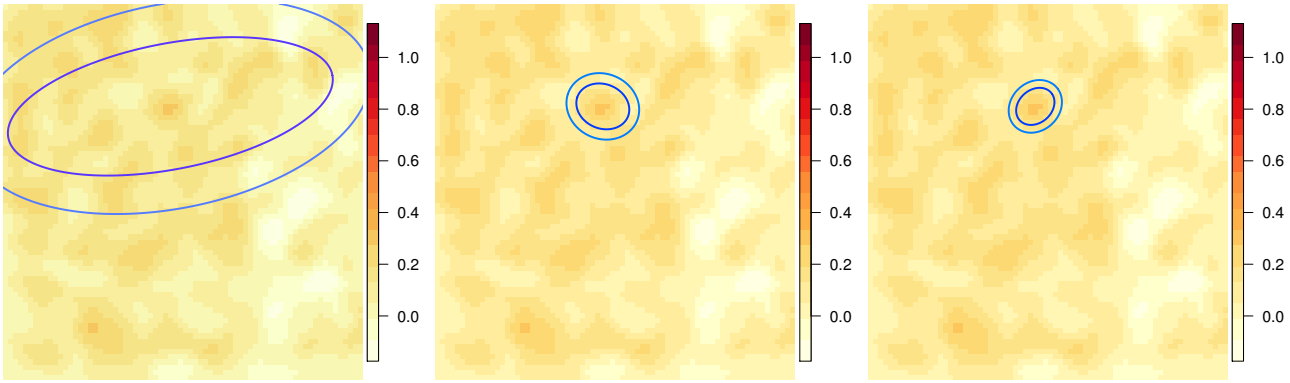
This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.



(a) Strong signal, $\lambda = 2$ (b) Strong signal, $\lambda = 100$ (c) Strong signal, $\lambda = 200$



(d) Weak signal, $\lambda = 2$ (e) Weak signal, $\lambda = 100$ (f) Weak signal, $\lambda = 200$



(g) Noise, $\lambda = 2$ (h) Noise, $\lambda = 100$ (i) Noise, $\lambda = 200$

Figure S-1: The effect of λ on optimal ellipses drawn on the illustrative images of Figures 1 and 2 having strong (top row, a–c), weak (middle row, d–f) and no true (bottom row, g–i) signal.

the optimal inner ellipse and the corresponding outer ellipse (using Step 3 of our algorithm) after optimizing $\mathcal{V}(a, b, \theta)$ for the three illustrative cases of Section 1.1 after processing with a matched filter, resulting in images as in Figure 2. For presentation clarity, the images in the top row with the strong true signal is drawn using a different scale than the other two sets of images. The figures in the first column all have larger ellipses than their corresponding counterparts in the next two columns which have ellipses in decreasing order of size. Thus, a larger regularization parameter results in

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

smaller ellipses, because of the heavier penalty put on noise pixels (which have negative $C_\sigma(\tau(u,v))$ values). The differences in the sizes of the inner ellipses, however, decrease for larger values of λ (last two columns).

S-2.2 Diagnostic Checks for Model Assumptions in Vibrothermography Specimens

We report results on some checks to evaluate the assumption of normality in the NIM model for

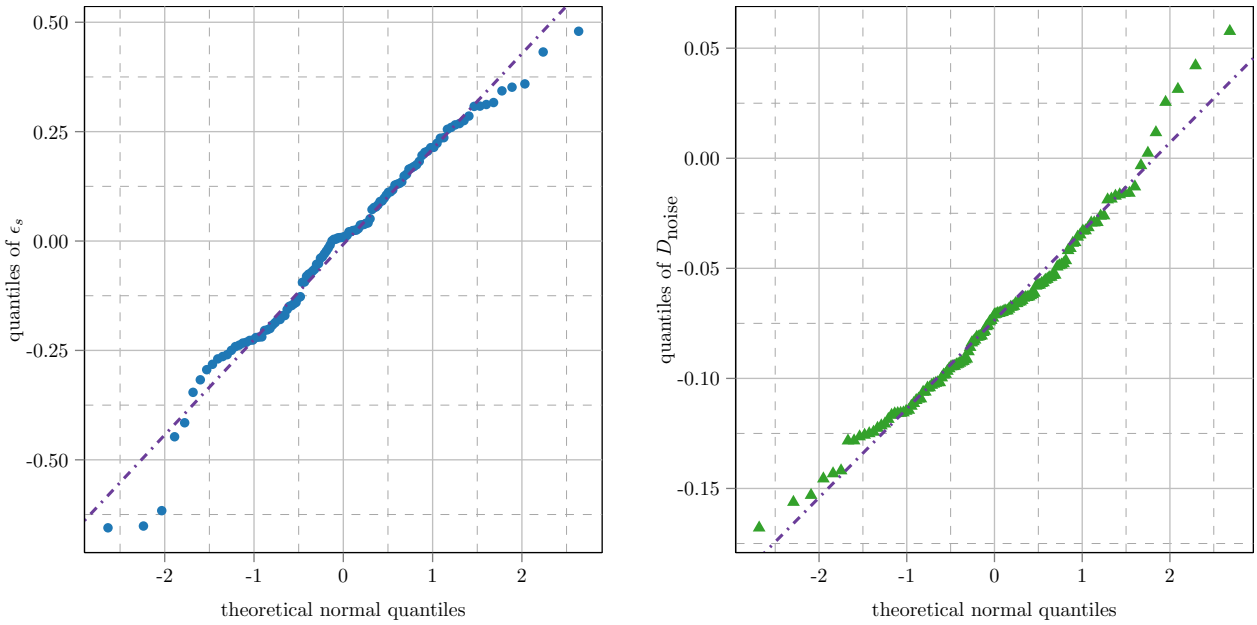


Figure S-2: Quantile plots for evaluating the assumption of normality in the NIM model for the (a) flawed and (b) flawless specimens.

the vibrothermography specimen images. Figure S-2 provides quantile-quantile plots for the residuals obtained upon fitting the NIM to flawed and flawless specimens. (Note that since there is no signal in a flawless specimen, the residuals are essentially the same as D_{noise}.) These plots indicate reasonably good agreement with the normal distribution. A formal test for normality using the Shapiro and Wilk (1965) approach yielded p-values of 0.109 and 0.07 for flawed and flawless specimens, providing support for the normality assumption. (We recall that in most real applications, the Shapiro-Wilk test will tend to find departures from a normal distribution with moderately large to large sample sizes.) We conclude by noting that we are not making inferences on the tails of the distribution and so the procedure is expected to be robust to moderate departures from the normality assumption in the NIM model.

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

S-3 Performance Evaluations – Supplement

S-3.1 Derivation of the NIM Model

We have, from the NIM model in Section 2.3.2

$$D_{\text{obs}} = \max(D_{\text{signal}}, D_{\text{noise}})$$

where

$$D_{\text{signal}} = \beta_0 + \beta_1 \log_{10}((\text{flaw size})) + \epsilon_s,$$

with $\epsilon_s \sim N(0, \sigma_s^2)$ distributed independently of $D_{\text{noise}} \sim N(0, \sigma_N^2)$. Then the POD is given by,

$$\begin{aligned}
\text{POD(flaw)} &= \Pr(D_{\text{obs}} > 0) \\
&= \Pr(\max(D_{\text{signal}}, D_{\text{noise}}) > 0) \\
&= 1 - \Pr(\max(D_{\text{signal}}, D_{\text{noise}}) \leq 0) \\
&= 1 - \Phi\left[-\frac{\beta_0 + \beta_1 \log_{10}(\text{flaw})}{\sigma_S}\right] \Phi\left(-\frac{\mu_N}{\sigma_N}\right).
\end{aligned}$$

A similar argument holds for the POD(flaw) of the "PeakAmp" method, which is derived in Equation (3) of Li and Meeker (2009). To elucidate, since the peak intensity $\check{Z}$ of the raw unprocessed hottest frame image (in $\log_{10}$) is used to characterize the image, we have that a flaw is detected if $\log_{10} \check{Z}_{\text{obs}}$ is greater than some threshold given by $\log Z_{\text{th}}$. Then, from the NIM, we have $\log_{10} \check{Z}_{\text{obs}} = \max(Y_{\text{signal}}, Y_{\text{noise}})$ so that

$$\begin{aligned}
\text{POD(flaw)} &= \Pr[Z_{\text{obs}} > Z_{\text{th}}] \\
&= \Pr[\log_{10}(Z_{\text{obs}}) > \log_{10}(Z_{\text{th}})] \\
&= \Pr[\max(Y_{\text{signal}}, T_{\text{noise}}) > \log_{10}(Z_{\text{th}})] \\
&= 1 - \Pr[\max(Y_{\text{signal}}, Y_{\text{noise}}) \leq \log_{10}(Z_{\text{th}})] \\
&= 1 - \Phi\left[\frac{\log_{10} Z_{\text{th}} - \gamma_0 - \gamma_1 \log_{10}(\text{flaw})}{\kappa_S}\right] \Phi\left(\frac{\log_{10} Z_{\text{th}} - \nu_N}{\kappa_N}\right).
\end{aligned}$$

where $(\gamma_0, \gamma_1, \kappa_S, \nu_N, \kappa_N)$ are defined as in the last paragraph of Section 3.1.1.

References

Li, M. and Meeker, W. Q. (2009), "A Noise Interference Model for Estimating Probability of Detection for Nondestructive Evaluations," *Review of Quantitative Nondestructive Evaluation*, 28, 1769–1776.

This is an Accepted Manuscript of an article published by Taylor & Francis in Technometrics on February 18, 2016, available online : http://www.tandfonline.com/doi/abs/10.1080/00401706.2016.1153000.

Shapiro, S. S. and Wilk, M. B. (1965), "An analysis of variance test for normality (complete samples)," *Biometrika*, 52, 591–611.